Structured Invention Tasks to Prepare Students for Future Learning: Means, Mechanisms, and Cognitive Processes

Ido Roll

December 2009
CMU-HCII-09-105

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, Pennsylvania 15213

Thesis Committee:
Kenneth R. Koedinger (co-chair), Carnegie Mellon University
Vincent Aleven (co-chair), Carnegie Mellon University
David Klahr, Carnegie Mellon University
Dan Schwartz, Stanford University

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

This work was supported in part by the Pittsburgh Science of Learning Center which is funded by the National Science Foundation award number SBE-0354420, and by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B040063 to Carnegie Mellon University. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect those of the funding agencies

Abstract

Successful instruction should help students acquire robust knowledge and prepare them for future learning opportunities. However, we are yet to find a winning strategy for systematically achieving robust learning (Bransford & Schwartz, 2001). Accumulated evidence suggests that discovery learning does not help most students acquire the basic foundations, and direct instruction, on the other hand, often leads to a relatively rigid body of knowledge (c.f., Tobias & Duffy, 2009). Instructional technologies are in a similar pursuit of robust learning (Koedinger & Aleven, 2007). However, students working with discovery environments often do not receive adequate support and thus fail to achieve desired learning gains (De Jong & van Joolingen, 1998). Students working with intelligent tutoring systems receive appropriate support, but on tasks that may not prepare them enough to make sense of new situations.

Recently, Schwartz and colleagues devised a hybrid method called Invention as Preparation for Learning (IPL; Schwartz & Martin, 2004). In IPL students attempt to develop novel mathematical methods prior to (and not instead of) receiving direct instruction. While Schwartz and Martin (2004) showed that IPL is successful in preparing students for future learning, questions regarding the mechanisms and scalability of IPL remain largely unanswered.

This thesis focuses on understanding the sources of IPL's effectiveness, and using that to design technology that can scale up IPL. To address these issues, I conducted a series of classroom experiments to assess the effect of IPL on students' domain knowledge, motivation, and general invention skills, and to identify under what conditions and by what cognitive mechanisms IPL accelerates future learning; I contrasted different versions of IPL in order to identify its core components; and I created and evaluated the Invention Lab, a unique intelligent tutoring system for IPL.

This thesis makes contributions to cognitive science by better understanding the mechanisms and effects of inventions in learning. It contributes to the learning sciences by conducting comprehensive evaluations of a novel pedagogy. And it contributes to the field of human-computer interaction by designing, implementing, and evaluating a novel type of intelligent system, capable of adapting to users' knowledge in scientific inquiry tasks.

Keywords: Invention as Preparation for Learning; Robust Learning; Transfer; Scientific Inquiry; Intelligent Tutoring Systems; Direct Instruction; Constructivist Theory; Cognitive Tutor Authoring Tools.

Acknowledgements

I have so many things to be thankful for to so many people, but only have so many words. First and foremost I want to thank my committee members for their never-ending challenges, patience, and support. I was lucky to have four friends and mentors I admire on my committee. Letting them mess with my mind has been a true pleasure. Ken, Vincent, David, and Dan, each of you had a profound influence on the way I perceive thinking, learning, science, and my role in all that. Thank you so much.

My best friends were my non-official committee. I would have never made it without 4 letter acronyms - PSLC, PIER, HCII, and of course, EdFed, that does not even do very well as a 4-letter acronym. Thank you, Ruth Wylie, Amy Ogan, Erin Walker, Ryan Baker, Cristen Torrey, Elsa Golden, Matt Easterday, April Galyardt, Ian Li, Rosta Farzan, Behrang Mohit, Peter Scupelli, Elida Lasky, Eriki, Silvush, Leila, and others. Thank you, Debby, for everything. I also want to thank Michael Bett, Jo Bodnar, Gail Kusbit, Queenie Kravitz, and Audrey Russo for their ongoing hard work and help. The CTAT team gave me an opportunity to build my dream system - what a treat! Thank you, Jonathan Sewall, Sandy Demi, Bo Chen, and Martin Van Velsen. Few other people have shown me the light when needed and had a tremendous influence on my thinking patterns and abilities. Especially, thank you Gautam Biswas, Roger Azevedo, Sharon Carver, John Anderson, Bryan Junker, and Marsha Lovett.

My teenager middle school experience back in the '80s was pretty good, and can only be matched by my current middle school experience. Thank you, Beth McCalister, Tom Kendro, Ed Wellman, Ed Vettel, and the other teachers. Especially, huge thanks for the middle school students I worked with. "I have learned a lot of Torah from my teachers, and from my friends more than my teachers, and from my students more than everyone" (Taanis 7a).

Sometimes, when I was working too much, I needed a reminder for what life, friendship, family, and the world is about. It all came in one package, called Pittsburgh Playback Theatre. I am SO happy I chose the right adopting family. Thanks till the end of time to my inspirations, Roni

Ostfield, Rich Keitel, Paulo Nzambi, Gentry Thurman, Demetria Marsh, Amy Guterson, Andrea Scheve, Jewel Barber, and Ashley Archambeau. I want to thank my amazing family, Ofira and Omer, for teaching me everything else. I have so many things to be thankful for to so many people; it will have to wait for another time.

This project is yet another attempt to bring worldwide peace. If you are taking the time to read these lines, please also take a minute to reflect on the role of schooling and learning in promoting peace between individuals, communities, and societies. Can we use middle school math to bring worldwide peace? Well, can we afford not to? We have the responsibility to help people become independent thinkers, aspire for more, and achieve that. We should help our students be receptive, respectful, and have the ability to adopt multiple perspectives. What is better than middle school math to achieve that?

To Kuja, Omi, and Sumsum, who reinvent me every moment.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

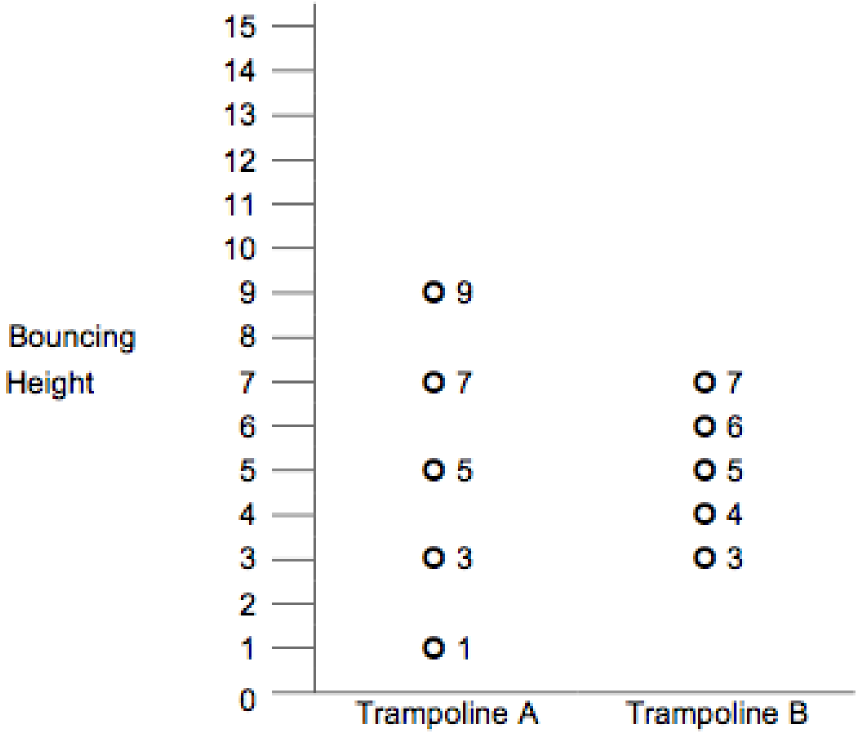
Rapid developments in today's world demand corresponding changes in the workforce and the world’s citizenry. To achieve personal, professional, and national growth, people should be life long learners and acquire necessary knowledge via on the job training. President Obama outlined his vision for education using the following words: "You’ll need the knowledge and problem-solving skills you learn in science and math to cure diseases like cancer and AIDS, and to develop new energy technologies and protect our environment. You’ll need the insights and critical thinking skills you gain in history and social studies to fight poverty and homelessness, crime and discrimination, and make our nation more fair and more free. You’ll need the creativity and ingenuity you develop in all your classes to build new companies ... ". In all these examples the president talked about using the knowledge acquired in school to further develop and learn, and not as a finite e product. More specifically, the president talked about three core sets of skills, all of which are at the focus of this thesis: problem solving skills, critical thinking skills, and creativity.

These characteristics of desired knowledge affects the goals of schooling. Rather than a rigid pre-defined body of knowledge, schools should help students achieve robust learning that will prepare them for future challenges (Collins & Halverson, 2009; Halpern, 1998). Such robust learning should include strong foundational knowledge and general learning skills, because both are required to transfer the learned knowledge to novel situations (Hatano & Inagaki, 1986; Koedinger & VanLehn, 2006). There is a widespread agreement that helping students transfer and expand their knowledge is an important goal of education (or, in the scope of this thesis, of math education; Schoenfeld, 1992). However, despite more than a century of studies (Judd, 1908; Thorndike & Woodworth, 1901), it is not yet clear how to systematically achieve that goal (Barnett & Ceci, 2002; Bransford, Brown & Cocking, 2000; Schwartz, Bransford & Sears, 2005). Most studies of transfer tend to focus on two alternative forms of instruction: direct instruction vs.

constructivist learning (Kirschner, Sweller & Clark, 2006; Tobias & Duffy, 2009). By direct instruction I refer to explicitly proving students with all the information regarding the target learning goals (Kirschner, Sweller & Clark, 2006). By discovery learning I refer to giving students the responsibility to infer the underlying models that drive observed phenomena, possibly while giving them assistance at the process level (de Jong & van Joolingen, 1998). Recently, more voices have started to argue for some combination of the two strategies (Bransford & Schwartz, 2001; Koedinger & Aleven, 2007; Rittle-Johnson, 2004).

This thesis explores one of these hybrid approaches, termed Invention as Preparation for Learning (IPL, Schwartz & Martin, 2004). IPL is a teaching strategy that uses constructivist activities and direct instruction in a complementary fashion (Schwartz & Martin, 2004). First, students are asked to invent general methods (and their equivalent mathematical expressions) to evaluate a set of examples, or cases, with regard to the target concept. Students may or may not succeed in inventing a valid method; the challenge is intentionally designed so that students will be only partially successful, but will learn about some of the key features of the target domain. The knowledge that is acquired during invention activities prepares students to learn better from subsequent direct instruction (Schwartz & Bransford, 1998). Figure 1 shows an example of an invention task in the domain of statistics. In this example students are asked to invent a method for comparing the variability of two datasets. The cover story tells students to invent a method for determining which trampoline is more fair for the Olympic games, that is, more consistent. While the materials of the invention tasks are not unique in and of themselves (similar problems are commonly given as practice items, following instruction), the combination of materials, timing, scaffold, and directions, is unique to invention tasks. Regardless of whether students succeed, following the invention attempt, students receive direct instruction on canonical solutions for the same problem, and practice these. For example, the invention task described above is followed by direct instruction on Mean Absolute Deviation (the average distance from the mean) and corresponding practice.

The Bouncers Trampoline Company tests their trampolines by dropping a 100 lb. weight from 15 feet. They measure how many feet the weight bounces back into the air. They do several trials for each trampoline. Here are the results for two of their trampolines:



Create a method for determining which trampoline's data points are closer to a single point. You should use the same method to evaluate both trampolines. Your method should give a single value for each trampoline. Write your methods in steps so that other people can apply it.

Figure 1: An example of an invention task. This task is based on Schwartz and Martin (2004) and was used in all three studies discussed in this document. Students usually begin with range. Subsequent sets of contrasting cases keep a single range, to encourage students to notice other features (e.g., sample size).

Invention tasks use contrasting cases to direct students' attention to deep features of the domain. Rather than analyzing a single data set, as commonly done, Invention tasks ask students to compare two or more sets of data that vary with respect to a single deep feature. For example, the two sets in Figure 1 have the same average and sample size but differ in their range. The use of contrasts is known to improve encoding and transfer in different contexts. Gibson and Gibson (1955) showed that contrasts can direct attention to deep perceptual features. Gentner, Loewenstein & Thompson (2003) demonstrated that contrasting cases can also assist conceptual understanding. Comparisons of cases that emphasize a target feature, ceteris paribus, are commonly used also in legal argumentation (Aleven & Ashley, 1997). Trumpower and Fellus studied contrasting cases in Statistics, and found that students often analyze these successfully even in the absence of formal knowledge (Trumpower & Fellus, 2008). This informal evaluation of the contrasts serves as a baseline against which students evaluate their inventions (Schwartz, Sears & Chang, 2007).

Invention tasks differ from discovery learning tasks in that students are not asked to reveal an underlying model (de Jong & van Joolingen, 1998), but instead, to develop genuine novel procedures. Compared with conventional problem-solving tasks, invention tasks are not intended by themselves to yield substantial observable learning gains. Students who fail to invent valid methods may not demonstrate learning gains immediately following invention attempts. Rather, invention tasks are designed to help students learn better from subsequent instruction, as can be assessed using future learning measures (Bransford & Schwartz, 2001; Schwartz & Bransford, 1998). Thus, unlike "stand alone" constructivist tasks, the invention tasks given in IPL are not being judged by themselves, but rather, as part of a larger instructional process.

Schwartz and Martin (2004) found evidence that IPL instruction, i.e., invention tasks followed by direct instruction and practice, improves students' ability to learn independently when implemented either by researchers or teachers, compared with direct instruction and practice alone (Schwartz & Martin, 2004). This effect, termed Preparation for Future Learning (Schwartz & Martin, 2004) or Accelerated Future Learning (Koedinger & VanLehn, 2006), was measured by giving students a learning resource embedded in the post-test, followed by a test item that required understanding, mapping, and applying the newly-given information. The learning resource given to students was a solved example on material that builds upon, but extends beyond, the procedures learned in class. Invention tasks were found to boost performance on future learning assessments even though students failed to invent generally valid methods.

Other instructional interventions that share features with IPL were also shown to improve learning gains. Kapur found that students who struggle with ill-defined problems prior to receiving instruction are better positioned to learn from subsequent instruction, compared with students who received the instruction upfront (Kapur, 2008; Kapur & Lee, 2009). The instruction evaluated by Klahr and Nigam (2004) also shared many characteristics with IPL. For their pre-test, Klahr and Nigam asked participants to invent a procedure for comparing ramps. Following the invention task, students received contrasting cases, reasoned about them, and then were given direct instruction from the experimenter. Though titled "Direct Instruction", this instructional sequence shares many features with IPL instruction as described above, and was shown to lead to better learning compared with discovery learning alone.

In contrast to these studies demonstrating the effectiveness of IPL methods, several researchers did not find benefits for IPL instruction over direct instruction alone. Furthermore, in some cases direct instruction was shown to be better than IPL on isomorphic or even all measures (Belenky & Nokes, 2009; Matlen & Klahr, 2009). As for the invention materials, Rittle-Johnson and Star (2009) showed that while asking students to contrast multiple problems using the same method is beneficial for learning, focusing on contrasting multiple methods for the same problem has even greater benefits. The iterative invention process in IPL has a sequential evaluation of different methods for the same problem. However, Star and Rittle-Johnson (2009) found that using multiple methods sequentially, rather than in parallel, is least productive.

A clarification regarding terminology: IPL, as a term, is ambiguous. It describes an instructional manipulation (i.e., a sequence of activities that takes place in the classroom), learning outcomes (i.e., an experimental result that shows that students are more prepared to learn), and a mechanism (i.e., that invention activities prepare students to learn). In this document I refer to IPL only as an instructional manipulation, that is, a sequence of activities. Doing so maintains the spirit of IPL as described in Schwartz and Martin (2004). More specifically, I use IPL to refer to an instruction that includes invention activities, followed by direct instruction and practice. This does not suggest that students are able to invent, or that invention activities prepare students for learning. In fact, the invention activities and their outcomes are part of the research questions discussed in this thesis.

Similarity, I use the term "invention activities" to describe an instructional event in which students are asked to invent novel solutions to different problems. These problems, therefore, are being referred to as "invention tasks". This is not to suggest that students invent successfully, or even invent at all, just like the term problem solving does not suggest that students solve the problems successfully. This definition of invention activities is similar to the one used by Schwartz and Martin (2004).

The invention activities in IPL instruction are followed by direct instruction and practice. I refer to the coupling of direct instruction followed by practice as "show and practice". Show and practice includes direct instruction on relevant procedures and concepts, and demonstration of the procedures, followed by practice opportunities for students. Notably, show and practice can be part of IPL, but is mainly an instructional manipulation in its own right. Table 1 summarizes these definitions.

Table 1: Definitions and terminology

Instruction	Description
Invention activity	An instructional manipulation in which students are asked to develop new solutions to different problems ("invention tasks").
Show and practice	An instructional manipulation that includes direct instruction on procedures and concepts, and examples for applying the procedures, followed by practice opportunities for students.
IPL	An instructional sequence that includes invention activities followed by show and practice (Schwartz and Martin, 2004).

1.1 Motivation

Many educational philosophers have argued for the benefits of high-agency, experiential learning (e.g., Dewey, 1964; Papert, 1980; Piaget, 2009). In addition, many educational researchers have argued for the inherent value of practicing scientific methods in the classroom (Kuhn, 2007; Savery & Duffy, 1995; Scardamalia & Bereiter, 1994). However, so far, most assessments of these and similar instructional approaches failed to show the desired learning gains in comparison to mere direct instruction (Kirschner et al., 2006). Learning how to systematically replicate the effect of IPL could bring the ideas mentioned above to the classroom while improving students' learning.

Stellan Ohlsson once said that the human race survived not because it could run faster, but because it could learn from every experience. "We are perfect learning machines" (Ohlsson, personal communication). IPL suggests that some instructional manipulations have hidden outcomes. It shows that seemingly ineffective instruction can yield superior learning gains when assessed appropriately.

1.2 Research questions

The overall goal of this thesis is to unpack the IPL process and its outcomes. It focuses on the following questions:

Q1: Instructional elements: What is the IPL process?

The first research question this thesis deals with is the operational definition of the invention activities and their components. The task that students receive during invention activities is clearly defined in Schwartz and Martin (2004): to invent a single, general method for measuring a target property of given data. Schwartz and colleagues also examined various contextual factors. For example, Schwartz and Martin (2004) describe the safe IPL classroom culture in which students are encouraged to be creative and generative, with no cost for errors. Schwartz and Martin also describe the type of feedback that teachers are encouraged to give during IPL. Sears examined another task element, the interactivity of the process, and found that IPL is more effective with small groups than with individuals (Sears, 2006). However, detailed specifications of the IPL process itself are yet to be defined. For instance, though examples for IPL interactions are detailed, it is not yet clear what the stages of IPL are. One of the main goals of this thesis is to supply an operational definition of the IPL instruction. More specifically, this thesis defines and evaluates the combination of materials and task elements that consists IPL, thus identifying the critical components of IPL. This was done in two stages: First I identified the different stages of IPL in a design study (study 0). Then I evaluated their relative contribution in two in-vivo studies (studies 1 and 2).

Q2: What is the overall outcome of IPL?

The IPL process was shown to lead to significant learning gain from pre- to post-test in the domain of statistics. It was also shown to yield better performance on future learning measures compared with show-and-practice instruction (Schwartz & Martin, 2004). However, the overall effect of IPL on students' learning is yet to be assessed and compared to show-and-practice instruction, especially with regard to motivational and metacognitive outcomes.[1] This thesis fills this void by comparing the two approaches along several dimensions.

First, I compare the effect of IPL on the flexibility of students' knowledge. I follow a distinction made by McDaniel and Schlager (1990) between transfer problems that require the application of a learned strategy (near transfer problems) and transfer problems that require the generation of a new strategy (future learning problems). McDaniel and Schlager asked students to discover solutions to several water jug problems (i.e., how to use several jugs to measure a certain quantity of water). They found that while these discovery tasks improve students' performance on the new-strategy items, they have no effect on near transfer problems. Schwartz and Martin (2004) add a twist to these results. They found that IPL instruction improves students' ability to solve new-strategy problems as long as students are given a learning resource, whereas the same learning resource did not help show-and-practice students to solve new-strategy problems. To further investigate the effect of IPL on knowledge flexibility, I compare IPL to show-and-practice instruction on problems requiring different levels of knowledge flexibility. Schwartz and Martin (2004) showed that students who engage in invention activities are better able to learn

[1] Recently, and in parallel to the work described in this thesis, Schwartz and colleagues have compared the effect of IPL to Direct Instruction on a series of isomorphic and near-transfer measures. (Schwartz, personal communication).

new-strategy, but have not evaluated whether this improvement is homogeneous across different types of assessments (e.g., isomorphic or near transfer items). My hypothesis, as supported by McDaniel and Schlager (1990), is that students who engage in IPL will acquire more flexible knowledge and thus will demonstrate better performance on new-strategy items. At the same time they will not show better ability to use existing strategies in alternative contexts (near transfer items). Furthermore, following the findings of Schwartz and Martin (2004), I hypothesize that the effect of IPL will be mainly on encoding and using new-strategy instructions.

This thesis also evaluates the effect of IPL on students' motivation and interest levels. I do so using both self-report and direct behavioral measures. The question of motivation is of interest due to the distinct interaction style between teacher and students, and the very different classroom culture that IPL helps bring about, compared to more typical forms of classroom instruction. In most math lessons students are in search of a single, pre-defined correct answer, known to the teacher. In IPL, in contrast, there is more than one valid way to do things, and more than one correct answer. While some correct solutions are known (such as using Mean Absolute Deviation to measure variability), other novel valid solutions exist (and can be invented by students). The teacher is not an all-knower, and every attempt is valuable. Two contradictory hypotheses can be put forward with regard to students' liking of the IPL process, compared with traditional instruction. On one hand students are likely to enjoy IPL due to its novelty, high agency, and accepting classroom culture. On the other hand, students may perceive IPL as confusing, wasting their time, and may be discouraged by their failure to invent valid methods (Koedinger & Aleven, 2007). This contrast is especially interesting as it pertains to students with high test anxiety. These students are pressured the most by the prevailing one-correct-answer policy, and at the same time, may be confused the most by the change of rules that IPL represents. I hypothesize that students will have mixed reactions to the IPL process, leading to higher variability of their liking ratings.

Last, this thesis evaluates the effect of IPL on students' metacognitive knowledge and self-regulated learning skills (SRL). During their quest for inventing valid mathematical methods students practice different metacognitive and SRL skills. For example, students need to evaluate the progress they are making towards inventing valid methods and judge whether their methods are satisfactory. The set of metacognitive and SRL skills students practice during invention activities resembles the set of skills students practice during scientific inquiry (Kuhn & Pearsall, 2000). Following the common wisdom that practice makes perfect, I hypothesize that students will get better at the specific scientific reasoning skills they practice. This is not to suggest that students will acquire better SRL skills overall. Rather, for the limited scope of the study, I hypothesize that students will acquire better domain-independent invention skills applicable to isomorphic invention tasks.

Q3: Cognitive mechanisms: What knowledge is acquired during invention activities, and how does it transfer?

Invention tasks have two uncommon properties. First, they lead to a positive effect on learning even when students' inventions are not mathematically valid. Second, they have a positive effect on acquiring future knowledge components that were not practiced during the invention activities themselves.

One of the goals of this thesis is to propose and evaluate several potential mechanisms that can explain this effect. More specifically, I will attempt to characterize what knowledge is acquired during invention activities, how it interacts with the subsequent instruction, and the conditions under which it yields a positive effect on future learning measures.

Q4: IPL and technology: Can a computer tutor effectively facilitate the critical elements of IPL?

IPL cannot be adopted very easily in educational practice; it requires teachers to be trained on how to implement this pedagogy in their classrooms. The complexity and subtlety of IPL instruction may lead to an inadequate implementation, and thus may fail to achieve desired learning outcomes (Kirschner et al., 2006). An alternative approach to scalability may be to use technology (Aleven & Koedinger, 2002). Can similar (or better) results be achieved when facilitating IPL using a tutoring system? Intelligent tutoring systems are known to benefit students during conventional problem-solving tasks (Koedinger, Anderson, Hadley & Mark, 1997; Morgan & Ritter, 2002). However, supporting students in more open-ended inquiry environments poses novel challenges to technology, and so far has not demonstrated comparable results (van Joolingen, 1999; Veermans, de Jong & van Joolingen, 2000).

As part of this thesis, I designed, built and evaluated a novel tutoring system for IPL called The Invention Lab. The Invention Lab is a unique intelligent tutoring system for scientific inquiry tasks, built using the Cognitive Tutor Authoring Tools (Aleven, McLaren, Sewall & Koedinger, 2006). In addition to its contribution to scaling up IPL, the Invention Lab also allows researchers to run more tightly controlled studies with invention tasks, while doing within-class manipulations. I hypothesize that the Invention Lab, utilizing cognitive models at the domain and the scientific-inquiry levels, will be successful at facilitating IPL.

Q5: Generalizability: Does the effect of IPL hold with different researchers and populations?

This thesis addresses two aspects of generalization. Schwartz and Martin write that *"(IPL) studies used relatively small sample sizes and narrow demographics, and it is important to see if the results hold more broadly."* (Schwartz & Martin, 2004, pg. 169). This thesis investigates whether IPL is effective also when used with a population other than the one used in Schwartz and Martin (2004).[2] The thesis further evaluates whether IPL can be systematically replicated by a different set of researchers than those who created the method. I hypothesize that the answer to both questions is positive, that is, IPL can be systematically replicated by a different set of researchers (led by me) in a population that differs from the one used by Schwartz and Martin (2004).

[2] Recent studies suggest that IPL is effective with a wide variety of populations, whether in below-average schools in the US (Schwartz, personal communication), in India (Kapur, 2008), or in Singapore (Kapur & Lee, 2009).

The subsequent sections in this thesis detail the three studies that were done in this project: study 0 (a small scale design study aimed at identifying the IPL components); study 1 (a controlled classroom study aimed at measuring the overall effect); and study 2 (a controlled classroom evaluation of the Invention Lab). Table 2 outlines the mapping of the different research questions to studies. In addition, the thesis describes in depth the different hypothesized cognitive mechanisms that explain IPL and the core components of the Invention Lab.

Table 2: Mapping of studies to research questions

Research question:	**Study 0: small scale design study**	**Study 1: paper and pencil**	**Study 2: The Invention Lab**
Q1: Instructional elements	• Identify critical components	• Evaluate sufficiency of intuitive ranking	• Evaluate role of design
Q2: Overall outcomes		• Measure effect on domain knowledge and motivation • Compare effect of IPL to a variant of direct instruction	• Measure effect on domain knowledge, metacognitive knowledge, and motivation
Q3: Cognitive mechanisms	• Suggest mechanisms	• Evaluate subset of mechanisms	• Evaluate remaining mechanisms
Q4: IPL and technology			• Evaluate the Invention Lab
Q5: Generalizability		• Evaluate IPL using a different set of researchers. • Evaluate IPL with two different levels of students	• Evaluate IPL using a different set of researchers. • Evaluate IPL with two different levels of students

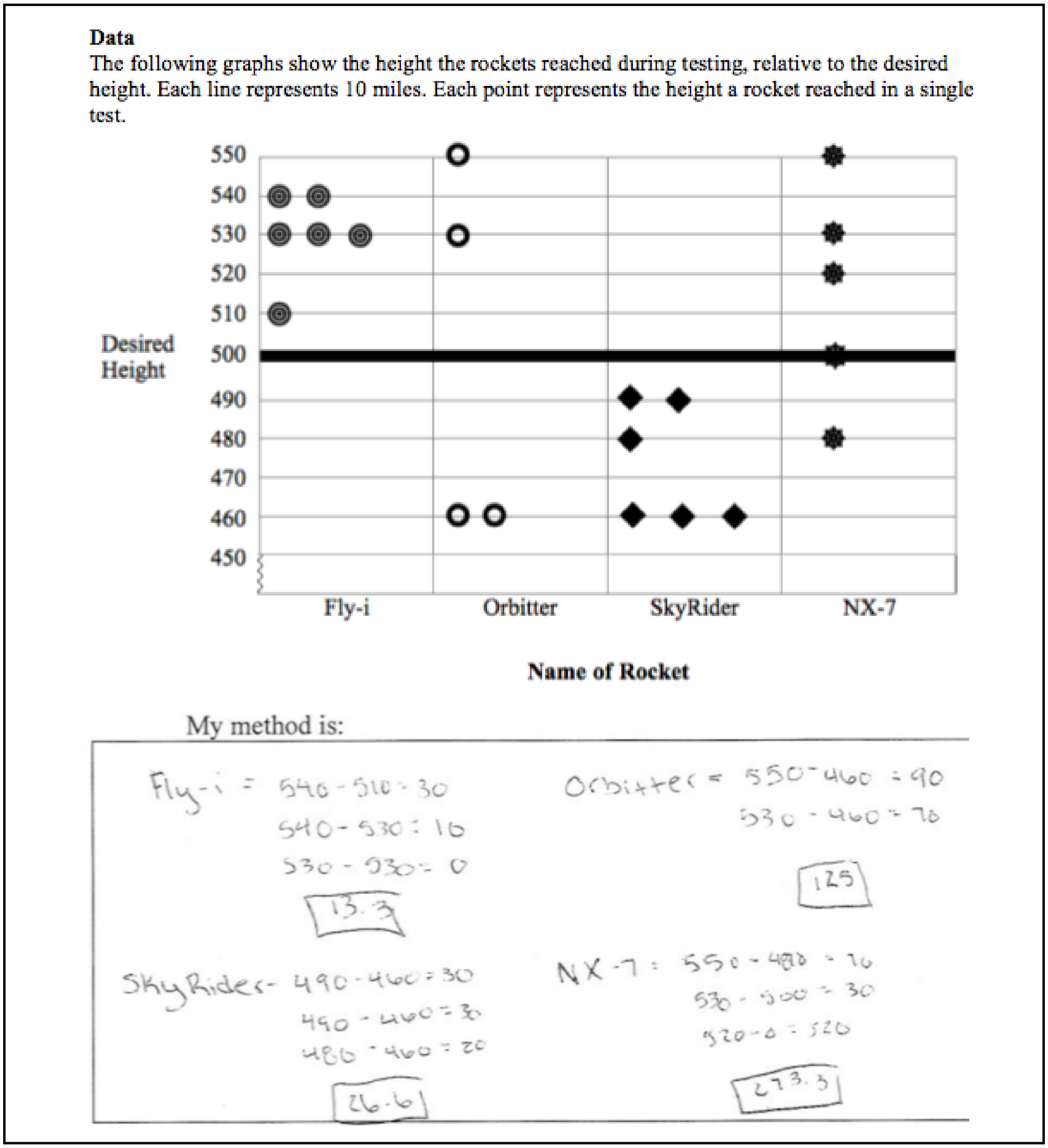
Chapter 2
Study 0: Identifying the components of IPL

2.1 Methods

The goals of study 0 were to experiment with the IPL process, map its components (Q1: Instructional elements), and prepare materials for studies 1 and 2. This study is an instance of the design research methodology (Barab, 2005; Brown, 1992; Collins, Joseph & Bielaczyc, 2003). During the study I taught an elective math class for one weekly period over one semester at a private school in Pittsburgh. Four students participated in this class (two boys and two girls), grades 6-8. The two boys were among the best in their classes and looked for extra challenges; the two girls were struggling and came to reinforce their mathematical knowledge.

During the study I experimented with different versions of IPL (e.g., individuals vs. pairs, with or without class discussion, with or without prompting students to use their observations to judge their methods, etc). The different elements were evaluated informally, by seeing how well the activity went and how much progress students made. The invention tasks covered a variety of topics (statistics, geometry, probability, etc).

Figure 1 has an example for a task that was adopted from Schwartz and Martin (2004), refined during study 0, and later used in controlled studies 1 and 2. Figure 2 shows a different invention task that was developed during study 0. The following example demonstrates how students interact with these tasks and how their thinking evolves. The example further demonstrates how these interactions contribute to the cognitive task analysis of the invention activities, and subsequently inform design decisions. Figure 2 shows an invention activity in the domain of variability, in which students were asked which NASA rocket is better for putting a satellite in orbit. This was not the first variability activity students did, and they had already identified few of the key features of variability - for example, that the formula should use all data points and not just a sample.



Data

The following graphs show the height the rockets reached during testing, relative to the desired height. Each line represents 10 miles. Each point represents the height a rocket reached in a single test.

My method is:

Fly-i = 540 - 510 = 30
540 - 530 = 10
530 - 530 = 0
13.3

Orbitter = 550 - 460 = 90
530 - 460 = 70
125

SkyRider - 490 - 460 = 30
490 - 460 = 30
480 - 460 = 20
26.6

NX-7 = 550 - 480 = 70
530 - 500 = 30
520 - 0 = 520
273.3

Figure 2: The data for the NASA problem and an example invention. In this problem students are asked to develop a method for calculating the “consistency” of rockets, which corresponds to the variability in the height they reach. The demonstrated method invented by a student finds the average distance between pairs of numbers, starting with the range and moving inwards to the next-furthest-apart pair of numbers, and so forth. Similar methods (that use recursive ‘ranges’) were common throughout studies 1 and 2. Notice that the student expressed the instantiations of the general governing rule, rather than the rule itself. Also, the student chose to write the method step by step (find pair wise distances one by one, and then take their average) rather than write it all at once using parenthesis. The student also makes some implicit actions, such as finding the average of the distances. Last, the method is under defined, since it does not define what to do in the case of odd number of numbers. In this case, the student chose to include a zero (see NX-7).

At first, one student added up the distances between each adjacent pair of numbers, which amounted to calculating the range of the given set of numbers. While range may be a suitable way to predict variability of some contrasting cases, it fails to discriminate between cases with identical range – for example, Fly-i and Sky Rider (as seen in Figure 2). Therefore, the student attempted to find a way to add up the distances without reaching range (for example, adding twice the distance by going up and down, or adding the distances as percentages of the largest distance), and in all cases he created an equivalent of the range formula. He concluded that adding up all distances was not the way to go, but had no explanation. During the discussion I tried to show that the distances cancel each other out (since (a1-a2)+(a2-a3)=(a1-a3)). Another student attempted to calculate all the distances between all possible pairs of data points. This method was found to work, in that when applied to the two contrasting cases, it conformed to the student's intuitions about the cases, but also to be too labor intensive. During the ensuing discussion, the students compared these approaches.

Upon resuming work, a third student thought that a selecting a fixed reference point could solve the problem. He suggested to use mid-range (that is, min+range/2), and estimated variability by calculating average distance from mid-range. (Note that this method is very close to a "real" formula for variability, the mean absolute deviation.) This invention led to additional interesting discussions that helped students understand the concept outliers and other general properties of data. In order to help students to understand the limitations of using mid-range as their reference point, I gave them the following contrasting cases: Rocket A: points scattered between 300 and 400, vs. Rocket B: ninety-nine points at 300 and one at 3,000. To me, it seemed intuitively clear that case B had lower variability. However, qualitative application of the mid-range method would suggest case A has lower variability. The student argued that indeed A was better. He argued that the single point at 3,000 was so far away, that the rocket should be punished, since satellites were very expensive[3]. Note that this was the only valid method defining

[3] This example is evidence of the large difference between the populations in study 0 and in studies 1 and 2. During study 1, in contrast to the example shown here, I found that most students did not know what NASA, satellites, and orbits were. As a result, this exercise was eliminated from study 2.

the notion of variability that I observed students create during the three studies. While the method itself is unique in that regard, the discussion pattern described above is typical of IPL.

2.2 Findings

These repeated experiences seemed to highlight three main stages of the invention activity: intuitive ranking, design, and evaluation (see Table 3). At the beginning of the invention activity students rank the contrasting cases intuitively according to the target construct (e.g., variability). This stage requires more than just intuition; it requires students to understand the target concept and its relevance to the contrasting cases. However, students at this stage do not have the required mathematical knowledge to make accurate quantitative observations regarding the contrasting cases. Students are then asked to design mathematical methods to measure the same construct. Last, students evaluate their methods by comparing their outcomes to students' initial ranking. Naturally, the last two components are iterative, and often are hard to tell apart. These three stages of invention are followed by a short class discussion, in which students share their inventions. Following the discussion students receive direct instruction and practice the learned content.

Table 3: Stages of IPL

Task Element		**Example**
Invention:	Intuitive ranking	Which trampoline seems more consistent?
	Design	Create a method for calculating the consistency of each trampoline.
	Evaluation	Does your method give the same ranking as your initial observation?
	Short class discussion	What methods did you try?
Show and practice:	Show	One common method that mathematicians use is Mean Absolute Deviation. Here is how to use it...
	Practice	Apply the Mean Absolute Deviation formula to the following problems: ...

Notably, the steps of invention match the common scientific method. The hypothetico-deductive method, as suggested by Whewell (1989) and advocated by Popper (2002), includes the following steps: Collecting data and making observations, raising hypothesis, identifying implications and predictions made by the hypothesis, and comparing these refutable predictions to the initial observations. The invention task, as was evident in study 0, follows a very similar structure (see Table 3). Therefore, the invention activity helps students practice an important set of skills. In addition, the scientific method, when supported appropriately, was shown to transfer well across domains and tasks (Chen & Klahr, 2008).

Table 4: IPL vs. the hypothetico-deductive scientific model

Invention stage:	Explanation:	Corresponding stage in the scientific method:
Intuitive ranking	Students compare the alternatives in the contrasting cases and identify the correct ranking of cases	Collecting data and making observations
Design	Students create a mathematical method, or model, that explains the observations made during the previous stage	Raising hypothesis; Identifying implications and predictions made by the hypothesis
Evaluation	Students evaluate their methods by comparing their outputs to their intuitive ranking	Comparing the refutable predictions made by the hypothesis to the initial observations

While every invention task is a form of scientific inquiry, not every scientific inquiry is a form of invention. Here are some of the unique characteristics of invention.

Intuitive ranking:

By asking students to compare, contrast, and rank cases that differ along their deep features, the invention activity directs students' attention to these features (for example, the contrasting cases given in Figure 1 emphasize spread). This is true especially when these features have low salience. Students are able to (intuitively) rank cases successfully even when discussing complex constructs such as variability (Schwartz & Martin, 2004; Trumpower & Fellus, 2008). However, students seem more confused when ranking contrasting cases in which variability competes with central tendency (e.g., 1, 4, 7 vs. 10, 11, 12 is more confusing than 1, 4, 7 vs. 3, 4, 5). This "conceptual Stroop effect" is not surprising, given that central tendency is much more common and familiar, and thus overshadows variability (Heckler, Kaminski & Sloutsky, 2008). Heckler et al. demonstrate that exposing students to contrasting cases in which the salient feature fails to explain the result (for example, it is held fixed with changing outcomes) help students encode the overshadowed features.

Design:

During the design phase students invent methods that should accurately measure the target construct. Students often under-define their methods. For example, students tend to skip steps, or to use vague terms (such as "count the points that are close together").[4] To encourage students to be more complete and specific, the task should emphasize the need to use formal mathematical notations (such as "count the points that fall within 20 units"). The importance of using mathematical language was demonstrated by Schwartz, Martin, and Pfaffman (2005), who asked students to reason verbally or mathematically about the balance beam problem (in which students are asked to predict which direction a balance beam would tilt once weights are placed on it). All students noticed the deep features of the balance beam domain - distance and weight. However, only students who reasoned mathematically were able to integrate the two dimensions into a single representation. Interestingly, students' thinking evolved even though their solutions were not complete, similar to the IPL effect.

Though the methods are mathematical, students in IPL feel most comfortable (and appear to have the least cognitive load) when describing the methods in steps, rather than a single formula (Heffernan & Koedinger, 1998).

Another way to encourage students to create coherent methods is to ask them to explain their methods to peers. When students prepare their methods to be understood by peers, they are more likely to validate that the methods are complete and well defined. Beyond insuring completeness, students may also benefit from preparation for teaching (Palincsar & Brown, 1984), whether or not they actually teach their method to their peers (Bargh & Schul, 1980).

A different aspect of the design process is students' understanding of the generalizability of mathematical methods. Most of them seem to exhibit some level of understanding of the generalizability of mathematical methods, in that they almost always apply the same invented method to all cases presented to them simultaneously, as a single set of contrasting cases. At the

[4] Observations detailed in this section were first made during study 0 and later reaffirmed during study 1 with a larger sample of students.

same time, students tend to design ad-hoc methods to solve each set of contrasting cases, without realizing that the methods should transfer between different sets of cases.

The fact that students tend to apply a single method to all contrasting cases within the same set also suggests that students have a single internal representation of their method. However, when writing down or talking about their methods, students almost always avoid talking about their method in abstract terms (even if specifically prompted to use abstraction) and instead instantiate their method right away with the different cases. Figure 2 illustrates a few of the solution patterns mentioned above.

It seems that design that is done in small teams (pairs or trios) is indeed more effective than individual design, as found by Sears with college level students (Sears, 2006). There seems to be a strong gender effect in this regard. In both the design experiment (study 0) and the in-vivo studies (studies 1 and 2), teams composed of boys had a higher tendency to break apart and work individually compared with teams composed of girls.

Evaluation:

The last stage of the invention activity is evaluation. During evaluation students use the contrasting cases and their observations to evaluate their methods. Tasks that support mapping between the mathematical problems and students' qualitative perception of the situation were shown to improve schema acquisition (Nathan, 1998). This form of self-assessment is called Situational Feedback (Nathan, Kintsch & Young, 1992). Mathan and Koedinger (2005) demonstrated a similar effect, showing that tasks that support self-detection of errors (and thus follow an intelligent-novice model) lead to superior learning gains.

When students' methods fail to generate a correct ranking for the contrasting cases, students are expected to debug and revise their methods. When the methods produce the desired ranking students move on, that is, attempt to apply these methods to new sets of contrasting cases.

Each invention activity (that is, a single cover story) takes about 30 minutes. The last segment of the invention activity is class presentation, in which students are asked to present their methods to the whole class. While the students in study 0 were thrilled to present their ideas, they did not seem to care much for their peers' inventions. It seems that under tight time constraints, achieving a productive peer critiquing process (e.g., White & Frederiksen, 1998) is very challenging. Instead, the class discussion may play a motivational role, by encouraging students to work harder. Assessing the necessity of the peer critiquing process and its cognitive and motivational benefits in the context of IPL is outside the scope of this thesis.

Studies 1 and 2 go on to evaluate the necessity of these stages. More specifically, study 1 compares intuitive ranking only with full invention, and study 2 compares intuitive ranking and evaluation (but no design) with the full invention. Given the distinct cognitive role of each phase, and tight correspondence between the overall process and the scientific method, I hypothesize that all three stages are necessary to achieve positive effect on learning.

2.3 Cognitive processes and acquired knowledge

Hypothesized mechanisms for the effect of IPL should address the following two questions: what potential knowledge components or dispositions are acquired during invention (even when invention itself fails)? And how do these transfer to future learning tasks? In addition, any suggested mechanism should make refutable predictions. Analysis of the IPL process suggests several such mechanisms (see Figure 3). These are not necessarily mutually exclusive. Studies 1 and 2 test the predictions that can be derived from these hypotheses.

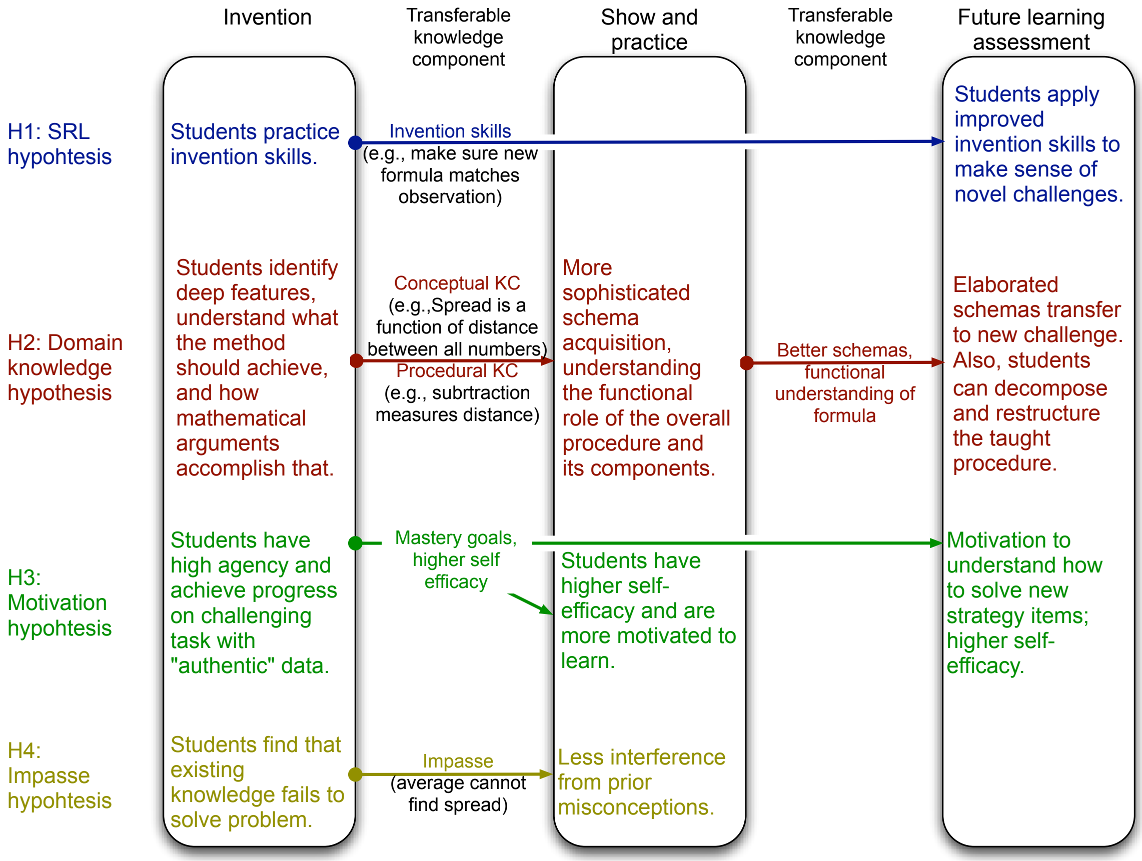

Invention
Transferable knowledge component
Show and practice
Transferable knowledge component
Future learning assessment
H1: SRL hypohtesis
Students practice invention skills.
Invention skills
(e.g., make sure new formula matches observation)
Students apply improved invention skills to make sense of novel challenges.
H2: Domain knowledge hypothesis
Students identify deep features, understand what the method should achieve, and how mathematical arguments accomplish that.
Conceptual KC
(e.g.,Spread is a function of distance between all numbers)
Procedural KC
(e.g., subrtraction measures distance)
More sophisticated schema acquisition, understanding the functional role of the overall procedure and its components.
Better schemas, functional understanding of formula
Elaborated schemas transfer to new challenge. Also, students can decompose and restructure the taught procedure.
H3: Motivation hypothesis
Students have high agency and achieve progress on challenging task with "authentic" data.
Mastery goals, higher self efficacy
Students have higher self-efficacy and are more motivated to learn.
Motivation to understand how to solve new strategy items; higher self-efficacy.
H4: Impasse hypothesis
Students find that existing knowledge fails to solve problem.
Impasse
(average cannot find spread)
Less interference from prior misconceptions.

Figure 3: Hypothesized mechanisms that explain the effect of IPL. The three rectangles represent the three consecutive learning events (invention, show and practice, assessment). The arrows demonstrate what knowledge components could transfer from each event to the subsequent ones, according to each of the main 4 hypotheses.

H1: Self-regulated learning hypothesis

According to the self-regulated learning (SRL) hypothesis, during invention students acquire scientific reasoning skills. These skills help them later make sense of future learning assessment items. In short, this hypothesis suggests that students who invent during practice are better prepared to invent during the test. For example, during invention students may realize that their invention methods should make sense and match their intuitive ranking. Later, during assessment, IPL students may apply this monitoring behavior and other metacognitive skills in order to make sense of items that require new strategies. This resembles the finding of McDaniel and Schlager regarding discovery learning: "Requiring discovery of a strategy while in training encourages the activation or refinement of procedures that are useful for generating a novel strategy... (and thus) facilitate transfer to tasks requiring a novel strategy." (McDaniel & Schlager, 1990 pg. 129).

The SRL hypothesis makes two predictions. First, since the benefits of invention are domain independent, it suggests that IPL students will perform better on invention tasks in an unrelated domain compared with students who did not learn using IPL instruction. Second, since IPL students are better equipped to deal with novel invention tasks, they will be more likely to attempt items that require novel strategies (even if they do not reach successful completion). In other words, according to the SRL hypothesis, IPL students will have weak methods and corresponding dispositions that will encourage them to attempt challenges even in the absence of sufficient domain knowledge.

H2: Domain knowledge hypothesis

The domain knowledge hypothesis suggests several ways in which invention attempts, even if unsuccessful, can contribute to domain knowledge. The knowledge acquired during invention prepares students to learn better from the show-and-practice phase, and subsequently, perform better during assessment.

Students who invent are exposed to many features of the domain by virtue of attempting to invent general valid methods. For example, when attempting to invent a procedure for computing variability using contrasting cases, students may realize that variability is a function of the distance between all numbers, or may better understand the relationship between variability and central tendency. The better schemas acquired by IPL students can lead to better transfer of class instruction to future learning items (Judd, 1908). For example, when facing an assessment item that requires the comparison of variability to central tendency (e.g., estimating the relative significance of variability), students can apply their improved schemas to adapt their knowledge (e.g., calculate the ratio between mean absolute deviation and average).

During invention activities students also evaluate what mathematical procedures succeed (or fail) to capture these features. By doing so students are more likely to understand what functional roles the target procedure should include. For example, students may realize that the procedure should control for sample size, though not be able to find a mathematical way to do so. By setting these requirements form the target solution, a correct procedure that is taught later is not perceived as a set of arbitrary operations, but rather as a solution to a set of constraints on what a valid solution should achieve (Ohlsson, 1994). At times, students may be able to invent valid procedural components that satisfy these requirements (even if their overall solution is faulty). For example, students may realize that taking the absolute value of subtraction is a good measure of distance. Functional mental models were previously shown to lead to more flexible knowledge (Kieras & Bovair, 1984). Hatano and Inagaki (1986) describe a similar process in which explaining empirical knowledge using procedures can lead to adaptive expertise. Hatano and Inagaki describe three requirements for this process to take place: One, the learner should ask herself why the procedure achieves the results it does. Two, the learner should based their reasoning on data that was collected while attempting to apply the procedure to examples varied along their deep features. Three, the conceptual knowledge should be grounded in a basic model, often acquired perceptually.

The process described by Hatano and Inagaki explains in what ways functional procedural knowledge can lead to better integration of conceptual knowledge: integration between the different features, and integration with prior knowledge and experiences. A similar result was described by Schwartz and colleagues who found that reasoning mathematically about the balance-beam problem leads to more coherent knowledge (Schwartz et al., 2005).

The Domain knowledge hypothesis suggests that invention activities share many properties and outcomes with prompt self-explanations (c.f., Chi, De Leeuw, Chiu & LaVancher, 1994; Siegler, 2002). In both processes students search for an explanation, which encourages deeper processing and thus greater conceptual understanding. Notably, self-explanation was shown to improve learning also when students reason about their faulty solutions (e.g., Siegler, 2002). The invention task facilitates this process by making thinking visible, that is, by giving students tools to explain their methods (Anderson, Corbett, Koedinger & Pelletier, 1995).

The Domain knowledge hypothesis suggests that IPL students are better at decomposing the procedures learned in class and restructuring their components to construct solutions to future learning items, in a similar fashion to the transfer between text editors as found by Singley and Anderson (1989). Likewise, this hypothesis suggests that IPL students would do better at debugging procedures that fail on one of these components. In addition, at the conceptual level, this hypothesis suggests that we can identify direct mapping between features revealed during invention and features required to solve new-strategy items during assessment.

H3: Motivational hypothesis

The invention activities may have a positive effect on students' motivation. By letting students express and explore their own reasoning, without immediate judgment and negative feedback, students develop ownership of the problem and its solution process, which may increase motivation and lead to greater learning gains (Savery & Duffy, 1995). Several motivational factors may interact to achieve the IPL effect. Challenging tasks, on which students can make incremental progress and in which they have high agency, are known to increase self-efficacy (Paris & Paris, 2001). In addition, IPL was suggested to lead to an adoption, even if temporarily, of mastery goals over performance goals (Belenky & Nokes, 2009). These findings suggest that, following invention, students may be more motivated to learn and understand the given instruction. A similar explanation suggests higher motivation during future learning assessment.

This hypothesis predicts that IPL will have a positive effect on measures of self-efficacy and situational interest. In addition, it predicts a smaller effect on assessment on isomorphic items to practice, for which performance goals can yield high learning gains (Elliot, McGregor & Gable, 1999), and a larger effect on far transfer assessments (such as new-strategy items). Perhaps the stronger prediction it makes is that detected motivation will be positively correlated with performance on cognitive measures.

H4: Impasse prompted learning hypothesis

Invention tasks encourage students to apply their existing knowledge and evaluate its relevance to the current problem at hand. During this process, students realize that their naive beliefs (e.g., average can do everything with data) cannot solve the new challenge (e.g., measure spread). While the other hypothesis identify knowledge that is transferable from invention to show-and-practice and eventually assessment, this hypothesis suggests that what transfers is the realization that students lack the relevant knowledge. This realization can facilitate conceptual change (Nussbaum & Novick, 1982; Scott, Asoko & Driver, 1991). Heckler showed that merely exposing students to the fact that salient factors cannot explain a certain phenomena (e.g., average cannot explain spread) is sufficient to have them look for an alternative explanation (Heckler et al., 2008). By realizing that their existing knowledge does not suffice to solve the invention task students may also reach productive impasses, which may prepare them to learn from subsequent instruction (VanLehn, Siler, Murray, Yamauchi & Baggett, 2003). Siegler describes a similar behavior with the balance beam (Siegler, 1983). He explains that when students notice that their own naive rules make wrong predictions they become motivated to encode new rules.

This hypothesis makes two predictions, relevant to our studies. First, it suggests that instructional manipulations that help students realize the limitations of their prior knowledge will lead to better learning. Second, it suggests that the biggest effect will be on items isomorphic to the items on which students reached impasses.

Table 5 summarizes the hypothesis raised in Study 0.

Table 5: Falsifiable predictions made by the different hypotheses. These predictions are evaluated in studies 1 and 2, in order to identify what knowledge is acquired during invention. The hypotheses compare IPL instruction to show-and-practice (or other form of reduced IPL) instruction. These control conditions were used in studies 1 and 2.

Hypothesis	Predictions
H1: Self-regulated learning hypothesis	• IPL students are more likely to attempt new challenges. • IPL students perform better on invention tasks in a different domain.
H2: Motivation hypothesis	• IPL students are more motivated to learn (and are especially more likely to adopt mastery goals) • There is a significant correlation between motivational measures and learning outcomes.
H3: Domain knowledge hypothesis	• There is direct mapping between features identified by students during invention attempts and features required by assessment items that evaluate flexible knowledge. • IPL students are more capable of diagnosing errors in variations on procedures learned in class.
H4: Impasse hypothesis	• Students who reach an impasse during invention perform better during assessment. • Reaching an impasse has the largest effect on knowledge that directly resolves the impasse.

Chapter 3

Study 1: Cognitive and motivational effects of IPL

3.1 Focus

Study 1 addressed 4 of the 5 research questions posed above. With regard to Q1: Instructional Elements, the study evaluates whether the intuitive ranking phase by itself leads to robust learning. Will students who are engaged in the full invention process show superior learning compared to students who merely rank the cases prior to instruction? Or does IPL require no more than making informed observations using contrasting cases? While analyzing the cases and ranking them, students notice the deep features, and may even realize that common methods do not suffice. In addition, since the intuitive ranking stage is very short, these students may benefit from more time for direct instruction and practice.

With regard to Q2: Overall Outcomes, the study evaluates the effect of IPL on students' domain knowledge and motivation, and compares it to the effect of a variant on direct instruction.

The study also addresses Q3: Cognitive Mechanisms by evaluating many of the predictions detailed in Table 5. The remaining hypotheses will be evaluated in study 2.

The study was conducted by me in a public school in the Pittsburgh area. The school performs below average on the standardized tests, and its population is different from the one used by Schwartz and Martin (2004). This addresses Q5: Generalizability.

3.2 Methods

(a) Design

The study compared two conditions: *Full Invention* and *Ranking Only* (see Table 6). Students in both conditions received contrasting cases and were asked to rank them according to the target concept (intuitive ranking stage). The first topic was central tendency and graphing. In the second topic, variability, students were instructed to also evaluate whether Average works.

This phase was followed by a class discussion of the correct ranking. In the second topic, the failure of Average to capture spread was also discussed. All students also received direct instruction (procedural and conceptual, see details under the Materials section) and opportunities for practice. The two conditions differed with regard to the invention activity following the intuitive ranking:

Table 6: The IPL process and experimental conditions in study 1

Activity type:	**Example task:**	**Experimental conditions:**	
		Full Invention	**Ranking Only**
Invention:			
Intuitive ranking	“Rank the following trampolines according to their consistency”	✓	✓
Design	"Invent a general mathematical method that yields a similar ranking"	✓	
Evaluation	"Did your method and your prediction gave the same ranking?"	✓	
Class discussion	"What method did you use?"	✓	✓
Show and practice:			
Direct Instruction	“One method that mathematicians use is Mean Absolute Deviation…”	✓	✓
Practice	"Apply the canonical method to the following problems:"	✓	✓

Full Invention students were asked to invent mathematical methods for calculating variability. This process had two iterative stages, as described earlier: First, students designed general mathematical procedures or visual representations that, when applied to the cases, should yield rankings similar to their (intuitive) observations. Then, students evaluated their methods by comparing the rankings generated by these methods to their observations. When their methods produced the desired ranking, students moved on to the next set of contrasting cases (each problem included several sets of contrasting cases, emphasizing different features of the domain, such as range, number of points, central tendency vs. distribution, etc). A mismatch in the ranking led to an iterative debugging process, in which students attempted to identify the reason for the failure of their method and improve it. After approximately 30 minutes, students engaged in a short instructor-led class discussion prior to receiving direct instruction. This condition resembled the instruction tested by Schwartz and Martin (2004).

Ranking Only students received direct instruction immediately following the intuitive ranking stage and the class discussion. Since the intuitive ranking stage alone is much shorter than full invention, instruction given to Ranking Only students was more detailed, and included more opportunities for practice and feedback. The canonical procedure was demonstrated using the same contrasting cases and its outcomes were compared to students' initial observations. The Ranking Only condition resembled traditional direct instruction with the addition of a short, guided intuitive ranking activity using contrasting cases.

Since activities varied significantly between conditions, whole classes were assigned to one condition or the other (between-class design).

(b) Participants

The study took place in six 7th-grade classes at a public middle school in the Pittsburgh area (30% free lunch[5], 35% minorities). Three of the classes were regular classes and three were advanced (pre-Algebra classes). At both levels, two classes were randomly assigned to the Full Invention condition and one to the Ranking Only condition. In order to minimize the chances for selection bias I validated that the end-of-year and standardized-tests scores did not differ between classes. The study included two topics. Due to absentees, not all students participated in both topics. 96 students participated in the first topic (66 in Full Invention, 30 in Ranking Only, split rather evenly between regular and advanced classes). 78 students participated in the second topic of the study (45 in Full Invention, 33 in Ranking Only). Notably, more than half of the advanced students in the Full Invention condition missed the second topic due to an overlapping

[5] The rate of free-lunch is the percentage of students whose lunch is subsidized. Generally speaking, higher rate of free lunch corresponds to lower socioeconomic status and worse performing school.

activity (see Table 7). No significant correlation was found between attendance in the second topic and pre-test scores (the students who missed that day were the ones who had raised more money during a fundraising drive.)

Table 7: Participants in study 1. Half of the Full Invention advanced students missed the second half of the study.

		Central tendency and graphing	**Variability**
Full Invention:	Regular classes	28	26
	Advanced classes	38	19
Ranking Only:	Regular classes	14	14
	Advanced classes	16	19

(c) Procedure

The study spanned 4 days with two class periods (of 42 minutes each) per day (see Table 8). The first two days covered topics of central tendency and graphing. The subsequent two days were on variability. The instruction related to both topics followed a similar structure. Full Invention students completed the invention tasks on days 1 and 3, and had show and practice activities on days 2 and 4. Ranking Only students had show and practice activities on all four days. The "show" component of the instruction was delivered by me. Overall instructional time was identical in both conditions. On day 1, all students completed a pre-test on central tendency and graphing (no pre-test on variability was given under the assumption of a floor effect). Post-tests on each topic were administered at the end of the relevant practice on day 2 (central tendency and graphing post-test) and day 4 (variability post-test). Students completed a delayed post-test about a month after the study.

Table 8: Procedure of study 1. Ranking Only condition received about twice as long show and practice activities (in blue / dark gray) compared with IPL condition.

Day	Topic	IPL	Ranking Only
Day 1	Central tendency and graphing	Introduction & pre-test (20 min)	Introduction & pre-test (20 min)
		Invention task 1 (30 min)	Ranking & discussion (10 min)
			Instruction (50 min)
		Invention task 2 (30 min)	
Day 2		Instruction (40 min)	Instruction (20 min)
			Practice (35 min)
		Practice (15 min)	
		Graphing post-test (25 min)	Graphing post-test (25 min)
Day 3	Variability	Introduction (10 min)	Introduction (10 min)
		Invention task 3 (35 min)	Ranking & discussion (10 min)
			Instruction (50 min)
		Invention task 4 (35 min)	Practice (20 min)
Day 4		Instruction (30 min)	Instruction (20 min)
		Practice (15 min)	Practice (25 min)
		Variability post-test (35 min)	Variability post-test (35 min)
About day 32	Delayed post-test	Delayed test (15 min)	Delayed test (15 min)

(d) Materials

Learning activities

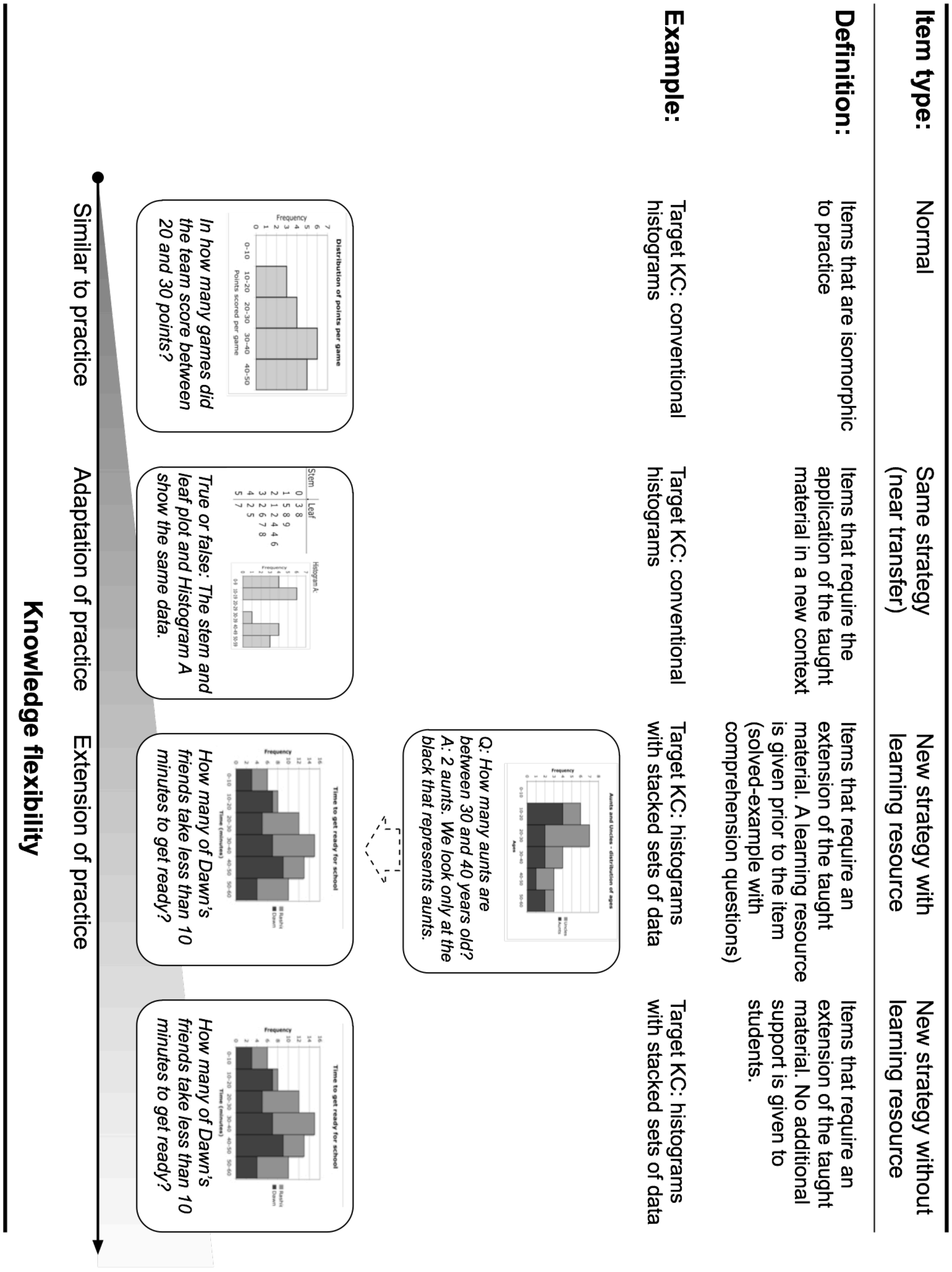
The study included two topics: (1) central tendency and graphing (histograms, stem and leaf plots, bar charts, box and whisker plots, mean, median, mode and range) and (2) variability (distribution, consistency, mean absolute deviation). For each of the topic, the instruction included two invention tasks with multiple sets of contrasting cases each. The two invention tasks for central tendency and graphing asked students to choose which class to attend (based on test scores) and which gender (boys or girls) shops more (based on revenue data). The two invention tasks for variability asked students to identify which trampoline is more consistent (based on hypothetical factory testing data, see Figure 1) and which rocket is more predictable (based on hypothetical NASA tests, see Figure 2). The contrasting cases were used in both conditions. All students encountered them during the intuitive ranking phase and the instruction phase. In addition, the Full Invention students used the contrasting cases as basis for design. The materials were based on the lessons learned during study 0 and, prior to study 1, were piloted in the lab and in another class from the same cohort in the school.

In addition to invention tasks, materials also included PowerPoint instruction that presented the invention tasks and included the "show" component of the instruction. During this component students learned to use the Mean Absolute Deviation procedure (which estimates variability to be the average distance from the average). The procedure was taught in steps, to match students' tendency to express procedures, and to avoid complex symbols and concepts (such as Sigma). Two versions of instruction were created, both of which covered the same material using the given contrasting cases. However, given the extra time available for instruction in the Ranking Only condition, the PowerPoint in the Ranking Only condition was more detailed and included more examples and opportunities for feedback. Both versions of the "show" component included formative assessments (where students were asked to vote on the right answer). Students received feedback on their performance, and the overall class performance was used to emphasize different components of instruction. Both versions were very detailed and had been rehearsed several times in order to ensure minimal modifications between classes. An observer from the research team was instructed to take notes of any deviations from the planned instruction. No major deviations were identified.

During practice, students received a booklet with a procedural and conceptual problems. Procedural problems asked students to apply the procedures learned in class. Conceptual problem asked students to choose the appropriate procedure or representation for specific challenges, and what conclusions can be reached by using specific procedures or representations. The practice materials were identical in both conditions, though, due to time constraint, Full Invention students completed many fewer items than the Ranking Only condition. None of the students finished all the practice items. Students received feedback on their final solutions from the instructor at the end of the practice session.

Assessment

To evaluate the effect of IPL in general and the intuitive ranking stage specifically on students' knowledge flexibility, the tests included items that required different levels of knowledge flexibility (see Figure 4 for example, and Appendix 3 for the full test). The first type of items was isomorphic to items students practiced in class, thus termed "isomorphic items". These test items had the same structure as problems solved during practice and just varied in terms of surface features. These changes were limited to changes in the cover stories and in the specific numeric values used. The second type, near transfer items, required the application of knowledge taught in class in a new context. While the knowledge learned in class was sufficient to address these questions, their structure was different from what was practiced in class. For example, during instruction students went back and forth between data tables and different representations. However, on the test, two near transfer item asked students to match up different representations of the same data without explicitly going through the data table.



Item type:	Normal	Same strategy (near transfer)	New strategy with learning resource	New strategy without learning resource
Definition:	Items that are isomorphic to practice	Items that require the application of the taught material in a new context	Items that require an extension of the taught material. A learning resource is given prior to the item (solved-example with comprehension questions)	Items that require an extension of the taught material. No additional support is given to students.
Example:	Target KC: conventional histograms	Target KC: conventional histograms	Target KC: histograms with stacked sets of data	Target KC: histograms with stacked sets of data

In how many games did the team score between 20 and 30 points?

True or false: The stem and leaf plot and Histogram A show the same data.

Q: How many aunts are between 30 and 40 years old? A: 2 aunts. We look only at the black that represents aunts.

How many of Dawn's friends take less than 10 minutes to get ready?

How many of Dawn's friends take less than 10 minutes to get ready?

Figure 4: Assessment of knowledge flexibility. Four types of assessment items required growing levels of flexibility.

The third type of items was structurally dissimilar from the practice items in substantial ways and required the generation of a new strategy during the test. These strategies built upon, but extended beyond, the material learned in class. For example, students learned in class how to interpret conventional histograms that represent a single set of data. A new-strategy item asked students to interpret histograms with two stacked sets of data. The items shown in Figure 4 can demonstrate the different type of knowledge required for each assessment type. Students who can repeat the procedure for interpreting histograms can answer isomorphic items, which ask them to read or create simple histograms. These students may have mastered the procedure, but may also have shallow understanding of representations (that is, representation may simply mean matching a number in a table to a number in a graph). Students who gained a more holistic understanding of data representations can solve near transfer items, which require them to translate one representation to another without going explicitly through data tables. While the structure of this task is different than interpreting representations, the required knowledge is identical – understanding what the different numbers and shapes mean in each representation.

New strategy items, on the other hand, require an extension of this knowledge. In order to solve the new strategy items seen in Figure 4, students should, among other things, (i) realize that multiple datasets can use a single graph (ii) learn how to read a legend and focus on the relevant dataset, (iii) realize that frequency corresponds to the size of the bar, rather than its highest value, and (iv) apply a procedure (e.g., counting or subtraction) to measure the size of the bar. These features of the required solution were not taught in class, and thus students who cannot extend their knowledge beyond what was taught would not be able to solve these items.

New-strategy items included two variants. One of the items included an embedded learning resource in the form of a solved example with comprehension questions. Items with embedded learning resource, termed future learning items (Bransford & Schwartz, 2001), evaluated students' ability to comprehend additional instruction and apply it to new-strategy problems. There were at least 3 other items (isomorphic or near-transfer) in between each learning resource (solved example) and the corresponding new-strategy item. The other new-strategy item on each test had no additional instruction, and thus evaluated students' ability to adapt their knowledge spontaneously. The graphing and central tendency test included 2 new-strategy items requiring different skills. Each test form had one new-strategy item with a learning resource and one without (counterbalanced between forms). The variability post-test included only a single new-strategy item, with a learning resource.

The Spelling Competition

Jerry participated in the school's spelling competition and achieved the following scores:
(9, 16, 8)

Jerry's math teacher asked his students to calculate the Mean Absolute Deviation of Jerry's scores.

Student B wrote the following:

- Step 1: I calculate the average of Jerry's scores. (9+16+8)/3 = 33/3 = 11
- Step 2: I subtract all the numbers from the average. 9-11 = (-2); 16-11=5; 8-11=(-3)
- Step 3: I find the average of these numbers. ((-2)+5+(-3)) /3 = 0/3 = 0. The MAD is 0.

a. Is this answer correct? Yes / Not
b. If not, in what step did Student B make a mistake? Step 1 / 2 / 3

Figure 5: Debugging items in the Variability post-test. The procedure suggested in this item is wrong since it does not use the absolute values of subtraction, thus including negative distances that cancel each other out (and therefore, by definition, will always give a variability of 0). All debugging items included variants on the taught procedure of MAD. The items presented the faulty methods in a similar structure to the one used in class (e.g., three distinct steps), and specified both the wrong method and its instantiation to the specific data.

The variability post-test also included near transfer items that tested students' ability to debug faulty procedures (see example in Figure 5). Students received 3 faulty methods, with similar surface features to the correct procedure learned in class (3 steps, similar wording, etc). For each of these methods, students were asked whether the method was correct, and if not, to identify the buggy step. Table 9 includes examples for all kinds of test items (in addition to the examples given in Figures 4 and 5). The complete tests are included in Appendix 3.

Table 9: Assessments of domain knowledge in study 1

Item type	Central tendency and graphing			Variability		
	Example	Pre?	Post?	Example	Pre?	Post?
Normal (isomorphic)	Make a stem-and-leaf plot of the following values: 41 14 38 28 39 28 47 66 23	✔	✔	Calculate the MAD for the following set of numbers: (5, 3, 8, 4, 5)	✔	✔
Near transfer	Here is the distribution of points the Jaguars Basketball Team scored last year in their games: (histogram) 1. What was more common – to score more than 40 point per game, between 30 and 40, or below 30?		✔	a. Can the average ever be negative? Please explain or give an example b. Can the MAD ever be negative? Please explain or give an example		✔
Debugging items				Jerry's math teacher asked his students to calculate the MAD of Jerry's scores. One student did the following: Step 1: Jerry had an average of (9+16+8)/3 = 33/3 = 11 Step 2: The distance from the average to all the points is: Between 9 and 11 is 2; between 16 and 11 is 5; and between 8 and 11 is 3. Step 3: I add up the three distances: (2+5+3)=10. The MAD is 10. a. Is this answer correct? b. If not, in what step did the student make a mistake?		✔
New-strategy with learning resource	*Learning resource:* There are a couple of ways to apply stem-and-leaf plots to numbers larger than 100. One of them is... *New strategy item:* The following table shows the NBA leaders for total 1. points scored during playoffs, places 6-18. (table). Make a stem-and-leaf plot of the data. (data is in thousands).		✔	*Learning resource:* Sometimes MAD is not the best way to compare the variability of two sets. The MAD is not a good measure when one set has very large values while the other set has low values... *New strategy item:* Four friends, who often play football together (but are not very good at it), decided to compare their performance to the top Steelers players. Here is the comparison of their overall receiving yards during the last season: (data table) It is clear that the Steelers had many, many more yards. Given the huge differences in their overall receiving yards, who would you say has lower variability? Show your calculations.		✔
New-strategy without learning-resource	See row above		✔			

Students' motivation and interest level were measured using several instruments. A behavioral measure evaluated voluntary effort, a generally accepted indicator of motivation (Schunk, Pintrich & Meece, 2008). Voluntary effort is often measured indirectly, by asking students to report their willingness to invest extra effort (e.g., Cordova & Lepper, 1996). Contrary to that, study 1 offered a direct behavioral evidence for voluntary effort. In five of the six classes in this study, the study took place in back-to-back math periods. To assess voluntary effort I counted how many students kept working on their tasks during breaks (3 minutes between back-to-back periods).

In addition, the assessment battery included a modified version of the Motivated Strategies for Learning Questionnaire (MSLQ). The MSLQ is a self-report questionnaire, previously validated and shown to capture students' motivation and use of cognitive and metacognitive strategies (Pintrich & De Groot, 1990; Pintrich, Smith, Garcia & Mckeachie, 1993). The original questionnaire assesses five scales of motivation and metacognition: self-efficacy, intrinsic value, math liking, self-regulation, and test anxiety. Each scale includes 3-5 items. The survey was adapted to the target population and task and piloted for comprehension. The adapted version included only 2-3 items per construct (see Appendix 4). The modified MSLQ was administered twice to all students - right before the pre-test and right before the final post-test (the questionnaires proceeded the tests in order to avoid any influence of perceived test performance). Test-anxiety items were included only in the pre-test questionnaire, to be used as a covariate.

Last, to evaluate the effect of IPL on students' engagement and situational interest, students were asked to compare the study with their everyday math class along several dimensions: enjoyment, challenge, effort, and perceived learning. These items were given only at the end of the test, and were incorporated into the modified MSLQ survey form. Unlike the MSLQ items, these items were not taken from an existing instrument, and were not independently validated (they were, however, piloted in advance using comparable population). Table 10 summarizes the assessments of motivation in study 1.

Table 10: Assessments of motivation in study 1

Assessment	Example	Pre?	Post?
Behavioral measure	The experimenter observed how many students kept working voluntarily during breaks	Throughout the study	
Modified MSLQ	I check that my answers make sense before I say I am done (Likert scale)	✔	✔
Situational interest	In the last several days I have learned more than I usually learn in math		✔
Test anxiety	I worry a lot about tests.	✔	

3.3 Results

Table 11 summarizes the results from study 1. There were no significant differences between the groups on the pre-test (Full Invention=33%, Ranking Only=36%, *F*(1,101)=9.7, *p*<.2). A repeated-measures ANOVA using the 6 identical test-items (with 23 problem steps) between the pre- and post-tests (with class level and condition as independent factors) showed significant learning during the study (*F*(1,91)=120.6, *p*<.0005, where students went from 35% to 53% correct).

Students' answers to the formative assessment questions during the instruction suggest that they did not reach mastery, and that longer instruction in the Ranking Only condition was not a waste of students’ time.

Table 11: Summary of results in study 1: score (SD).

Assessment		Regular classes		Advanced classes	
		Ranking Only	Full Invention	Ranking Only	Full Invention
T 1: Central tendency and graphing	Isomorphic	.43 (.13)	.50 (.21)	.73 (.13)	.68 (.15)
	Near transfer	.17 (.18)	.20 (.18)	.42 (.19)	.37 (.21)
	New strategy with learning resource	.07 (.27)	.07 (.26)	.50 (.52)	.58 (.50)
	New strategy without learning resource	.07 (.27)	.07 (.26)	.12 (.34)	.47 (.51)
Post test 2: Variability	Isomorphic	.69 (.27)	.48 (.36)	.83 (.27)	.84 (.20)
	Near transfer	.28 (.11)	.26 (.17)	.57 (.20)	.52 (.18)
	Debugging	.36 (.48)	.36 (.48)	.56 (.50)	.67 (.47)
	New strategy with learning resource	.00 (.00)	0.03 (.18)	.05 (.22)	.00 (.00)

While students in both class levels did not differ in pre-test, the gap between students in the advanced classes to those in the regular classes grew bigger as time progressed. A repeated measures analysis shows significant time x student-level interaction, *F*(2,132)=17, *p*<.0005; pair wise comparisons reveal that the gap between conditions at post test 1 (central tendency and graphing) was significantly larger than at pre-test, *F*(1,93)=30, *p*<.0005; the gap at post-test 2 (variability) was significantly larger than at pre-test (*F*(1,76)=24, *p*<.0005) and marginally significantly larger than at post-test 1 (*F*(1,67)=3.8, *p*<.07; see Figure 6).

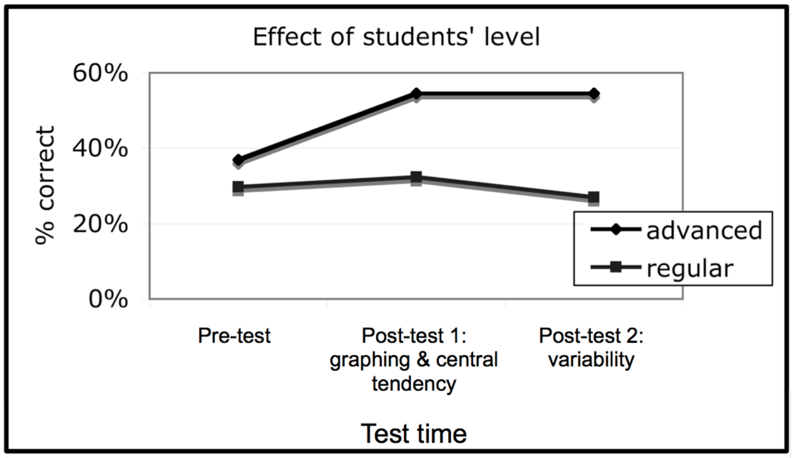


Figure 6: Effect of student level on test performance. Data shown above is overall score collapsed across conditions, and demonstrates that the students in the advanced classes were more likely to learn during the study (while there were no significant differences in the pre-tests of both groups). Post-test 1 and 2 were on different topics.

(a) Isomorphic and near transfer measures

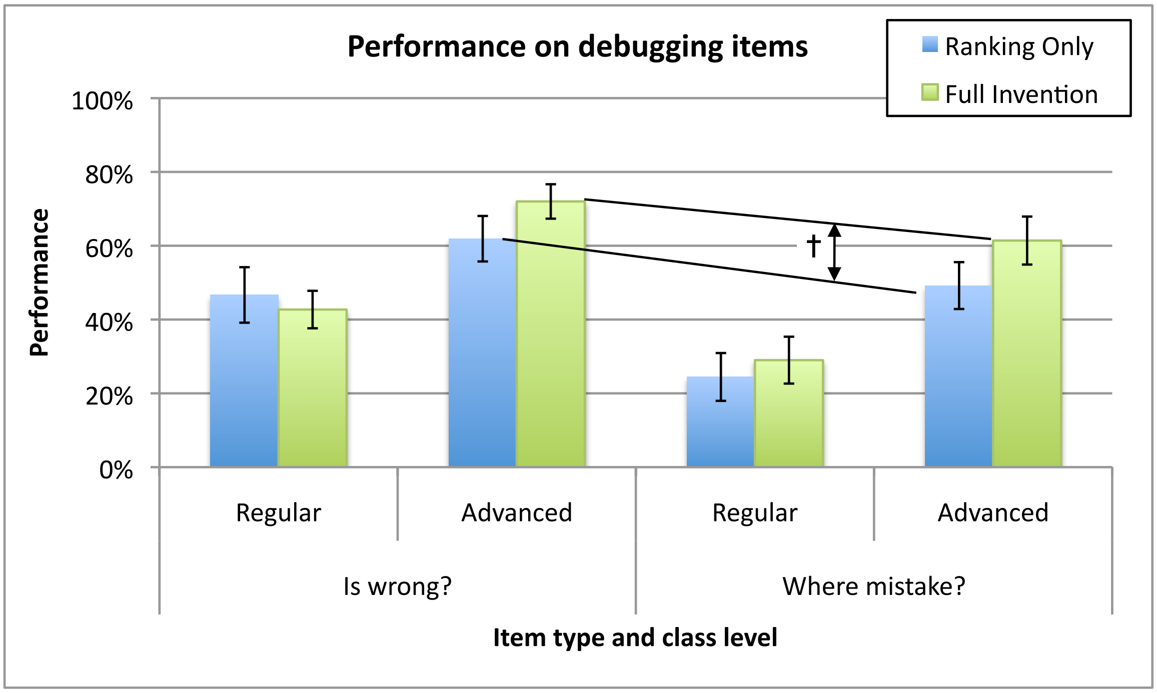
An ANCOVA of students' performance on isomorphic items on the graphing post-test (controlling for performance on pre-test) found no main effect for condition, but a significant interaction between condition and class-level (*F*(1,90)=4.1, *p*<.05). A separate ANCOVA for each class level showed that in the regular classes Full Invention students did marginally significantly better than Ranking Only students (50% vs. 43% respectively, *F*(1,38)=2.9, *p*<.1). There was no difference between conditions in the advanced classes (Full Invention: 68%, Ranking Only: 73%).

A similar analysis of student performance on the isomorphic items on the variability post-test showed a marginally significant interaction between condition and class level (*F*(1,73)=3.4, *p*<.07). Analysis within the levels found that in the regular classes Ranking Only students did marginally significantly better than Full Invention students (69% vs. 48%, *F*(1,37)=2.9, *p*<.1). There were no significant differences between the conditions in the advanced classes (Ranking Only: 83%; Full Invention: 84%).

Students in both conditions did equally well on near transfer items in both topics. On topic 1 both conditions scored 30% (*F*(1,90)=.2, *p*>.8). On topic 2 Full Invention students were 37%

correct and Ranking Only students were 46% correct (*F*(1,73)=.4, *p*>.5).[6]

The Variability post-test also included debugging transfer items. An ANOVA of students' performance in the advanced classes found a main effect for item type (*F*(1,76)=4.1, *p*<.05, with locating errors being harder than noticing errors, and a marginal effect for condition (Full Invention: 67%, Ranking Only: 56%. *F*(1,76)=3.8, *p*=.056, see Figure 7). The regular classes had a main effect for item type (*F*(1,90)=7.3, p<.01) but not for condition (Full Invention = Ranking Only = 36%).



(† - *p*<.1)

Figure 7: Scores on debugging items, Variability post-test. Full Invention students in the advanced classes performed marginally-significantly better than their Ranking Only counterparts on both types of items (*p*=.056).

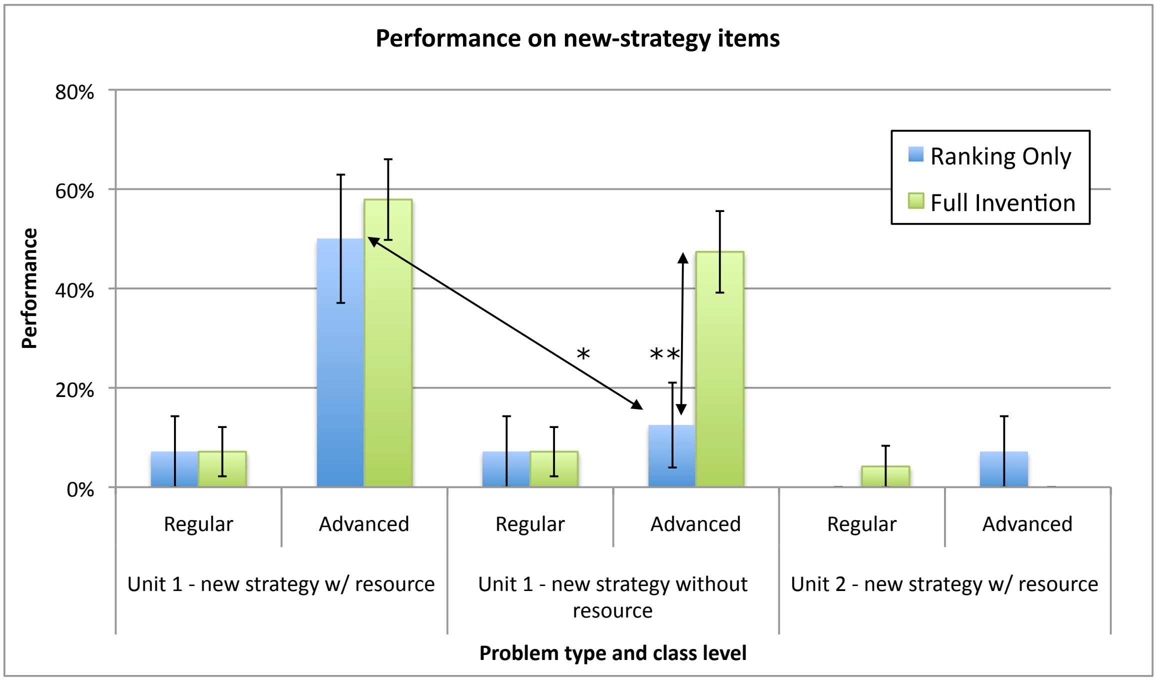
[6] The higher mean for Ranking Only students (45% vs. 36%) is a consequence of a higher ratio of advanced to regular students in the Ranking Only condition than in the Full Invention condition. See Table 11 for detailed scores by condition and class level.

There were no differences between conditions in the delayed post-test (Full Invention: 50%; Ranking Only: 48%; *F*(1,29)=.16, *p*=.7). The delayed test included isomorphic and near transfer items.

(b) New-strategy measures

The graphing post-test included new-strategy items with and without embedded learning resources. An ANCOVA of students' performance on new-strategy items without learning resources (controlling for performance at pre-test) found a significant advantage for Full Invention students (*F*(1,90)=5.3, *p*<.03; see Figure 7). There is also a significant interaction between condition and class level on these items (*F*(1,90)=3.8, *p*=.05). A separate ANCOVA for each class level reveals a significant effect only for advanced students (*F*(1,51)=7.9, *p*<.01; whereas Full Invention students scored 47%, compared with 12% in Ranking Only condition). Notably, the effect holds also when controlling for performance on isomorphic items on the same post-test (*F*(1,51)=6.5, *p*=.01). While Ranking Only students showed a significant drop in performance on new-strategy items in the absence of embedded instruction (*t*(15)=2.4, *p*<.03), the scores of Full Invention students on future learning items were not affected significantly by removing the learning resources (*t*(37)=1.0, *p*>.3). While Full Invention students in the advanced classes performed better on new-strategy items without resource, there was no difference in their attempt rate. Students in both conditions attempted on average 72% of the new-strategy items.

There was no difference between conditions in the regular classes (students in both conditions scored 7% on these items).



(* - *p*<.05; ** - *p*<.01)

Figure 8: Performance on new-strategy items. The variability post-test did not include new-strategy items without learning resource.

The variability post-test included only new-strategy items that followed embedded learning resources. In the variability post-test, scores on new-strategy items with learning resource were at floor (2% for Full Invention students, 3% for Ranking Only students). There was no significant effect for condition or its interactions on performance on these items. The variability post-test did not include new-strategy items without learning resource.

(c) Motivational measures

Behavioral measures

As explained earlier, voluntary effort was measured by counting how many students kept working during the 3 minutes breaks between back-to-back math periods. This was the case in 5 out of the 6 classes. On average, 16% of the students in the Full Invention condition remained in their seats to work during breaks, compared with 3% in the Ranking Only condition. Further analysis of the data for Full Invention students reveals a strong preference for invention tasks: An average of 24% of Full Invention students remained working when breaks interrupted invention tasks. In contrast, only 7% of them remained working when the breaks occurred during conventional practice (see Figure 9).

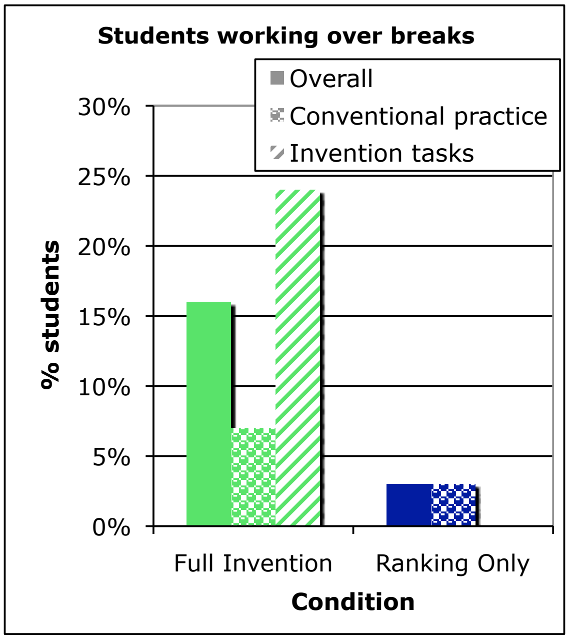


Figure 9: Voluntary effort during study 1. The graph shows percentage of students who kept working during breaks as a function of condition and activity.

MSLQ Questionnaire

The only significant difference from pre- to post-test on the MSLQ was an increase in the self-efficacy level of students in the Full Invention condition (*t*(43)=2.2, *p*<.03). While being statistically significant, this improvement is only modest, from 5.36 to 5.61, and does not hold when correcting for multiple comparisons using Bonferroni (see Table 12). There was no correlation between overall changes in the MSLQ responses (from pre-test to post-test) and performance on new-strategy items (controlling for class level and pre-test score, *r*=.18, n.s.).

Table 12: MSLQ results. The only significant difference from pre- to post- is a slight increase in the self-efficacy of Full Invention students. However, this increase is not significant when correcting for multiple comparisons using Bonferroni.

		Liking math	**Intrinsic value**	**Self regulation**	**Self-efficacy**
Ranking Only	Pre-test	4.16	4.71	4.58	5.45
	Post-test	4.15	4.44	4.59	5.38
Full Invention	Pre-test	4.44	4.92	4.39	5.36*
	Post-test	4.56	5.04	4.57	5.61

* - *p*<.05

Situational Interest

The situational interest questionnaire used a 1 to 7 Likert scale. The value 4 was used to describe a typical week, with values below 4 reflecting a decreased interest compared with a typical week, and values above 4 reflecting an increase in students' situational interest.

Students in the Full Invention conditions reported higher interest levels compared to students in the Ranking Only condition. While responses from Ranking Only students did not differ significantly from a typical week (*M*=4.3, *t*(30)=1.6, *p*=.12), responses from Full Invention students were significantly higher than a typical week (*M*=4.8, *t*(43)=5.0, *p*<.0005). There was no effect for level, gender, or learning gains. That is, students at all levels of achievement reported similar situational interest levels. Also, the variability in students' responses was similar across conditions (SD(full invention) = SD(Ranking Only) = 1.0).

Before the beginning of the study students reported their test-anxiety levels using the MSLQ instrument. An ANCOVA of situational interest with test anxiety as a covariate reveal a significant aptitude treatment interaction (see Figure 10, *F*(1,71)=5.0, *p*<.03), with a greater advantage for Full Invention for higher levels of test anxiety. There was no correlation between reports of situational interest and success on new-strategy items (controlling for class level and score on pre-test, *r*=.02, n.s.).

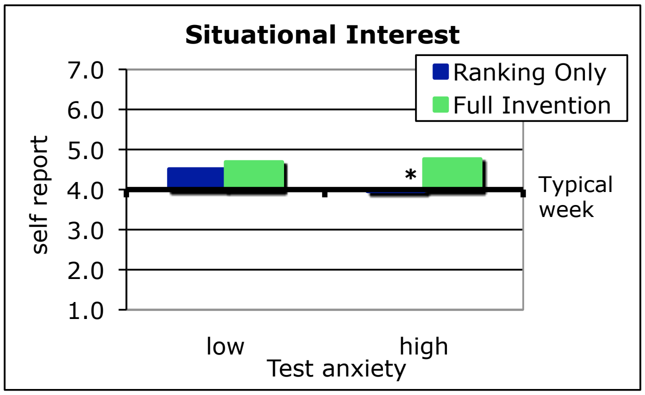


Figure 10: Situational interest as a function of test anxiety. The survey items asked students to compare the study to a typical week (4.0 on the Likert scale). Full Invention students (across levels) reported that the study was higher than a typical week in terms of required effort, interest levels, and perceived outcomes. The effect was especially significant for students who reported to have high test anxiety.

3.4 Discussion

(a) Summary of results

The results of study 1 inform the research questions in the following way:

Q1. Instructional elements

One of the goals of study 1 was to evaluate whether the intuitive ranking stage alone is sufficient to achieve the full effect of IPL. The results from study 1 provide evidence for benefits of the other stages. The complete invention cycle, including intuitive ranking, design, and evaluation, helped students make more sense of new-strategy items in the absence of learning resources, become more interested in the instructional activity, and possibly perform better on debugging tasks. Not all students demonstrated all effects, for instance, students in the regular classes did not exhibit superior performance on new-strategy or debugging items. Nevertheless, this study supports the idea that to the extent one wants to achieve better adaptability of knowledge, that it is worth engaging students in structured efforts to invent, in addition to studying intuition-enhancing cases. In other words, while engaging in design and evaluation may not be sufficient for achieving flexible knowledge, the results shown above suggest these stages are necessary steps in the process.

Q2. Overall effect:

Another goal of study 1 was to evaluate the effect of IPL on students' domain knowledge and motivation. Specifically, study 1 evaluated the hypothesis that IPL leads to increased flexibility in students' knowledge, and may achieve mixed effect on students' motivation. To evaluate this hypothesis, the study compared the performance of students in the Full Invention condition to that of students in the Ranking Only condition using a comprehensive set of measures. This comparison is relevant for the comparison of IPL to direct instruction, since the Ranking Only condition differed from direct instruction only in a short qualitative analysis of contrasting cases (and thus may do better than traditional direct instruction, and is unlikely to do any worse).

The results show different patterns in advanced and regular classes. In advanced classes, IPL led to improved performance on new-strategy items, and a marginally significant improvement on debugging items. There was no effect on isomorphic or near transfer measures. These findings echo the results reported by McDaniel and Schlager (1990). Notably, in the current study (study 1), these benefits for IPL were found even though none of the students invented a mathematically sound method during the invention phase as has also been observed by Schwartz & Martin (2004) and Kapur (2009).

IPL did not achieve similar results in the regular classes. In these classes there was no effect on new-strategy measures, and no clear pattern in isomorphic measures (marginally significant advantage for Full Invention on the first topic, marginally significant advantage for Ranking Only on the second topic). The performance of students in the regular classes on all six new-strategy items was at floor. Apparently, these items were too difficult for this sample of students, and thus the measure was insensitive to differences in flexibility of knowledge. To address this floor effect, study 2 uses new-strategy items that are within the ability of students in the regular classes.

Interestingly, even though IPL students had only approximately half the time for the show-

and-practice, they generally did as well on isomorphic measures (same for advanced, marginally better and worse for regular IPL students on topics 1 and 2, respectively). Looking at the bigger picture, it seems that students in both conditions did equally well on tasks for which they received some form of instruction - whether in class (on isomorphic and near transfer items) or embedded in the test (on new-strategy items with embedded learning resources). The cognitive effect of full IPL was mainly on new-strategy items with no instruction. This finding is at odds with earlier findings by Schwartz and Taylor (2004) who found that IPL improves students' ability to encode and apply future instruction but not solve novel problems without additional instruction. One explanation for the discrepancy between the studies is that the control group in Schwartz and Taylor (2004) was show-and-practice only, and these students did not engage in intuitive ranking of contrasting cases. Therefore, it may be that the intuitive ranking phase helped students in the current study to encode the novel instruction. After all, these cases (when used during show-and-practice) are essentially worked examples and greater use of worked examples has a powerful cognitive effective for novice learners (Salden, Aleven, Renkl & Schwonke, 2008; Sweller, van Merrienboer & Paas, 1998). An alternative explanation is that the embedded instruction and new-strategy items in this study may have been easier for students than the items given by Schwartz & Martin (2004) were for their students. Thus, advanced students in both conditions could solve the new-strategy items given adequate instruction, and in the case of IPL students, also in its absence.

The study also reveals effect of IPL on students' motivation for learning. Students in the Full Invention condition reported to have higher situational interest, and did more voluntary work during breaks. Furthermore, Full Invention students did more voluntary work during invention tasks than during conventional practice. The motivational benefits of IPL were seen mainly with students who have high test-anxiety. This suggests that incorporating IPL activities into everyday math classes may help change the attitude of these students towards learning math.

One of my hypotheses was that IPL may be a mixed blessing with regard to motivation, that is, while some students may enjoy it, others may find it frustrating. However, students' responses to the situational interest questionnaire had similar variability in both conditions, suggesting that IPL is not as controversial as hypothesized.

Q3. Cognitive mechanisms:

Study 1 gives us an opportunity to evaluate the different hypothesized mechanisms of IPL (see Table 13).

Table 13: Evaluating the predictions based on study 1. The results of study 1 inform the predictions made earlier in the following manner:

Hypothesis	Predictions	Evidence in study 1
H1: Self regulated learning hypothesis	• IPL students are more likely to attempt new challenges.	X
	• IPL students perform better on invention tasks in a different domain.	(not assessed)
H2: Motivation hypothesis	• IPL students are more motivated to learn (and are especially more likely to adopt mastery goals)	✓
	• There is a significant correlation between motivational measures and learning outcomes.	X
H3: Domain knowledge hypothesis	• There is direct mapping between features identified by students during invention attempts and features required by assessment items that evaluate flexible knowledge.	✓
	• IPL students are more capable of diagnosing errors in variations on procedures learned in class.	✓
H4: Impasse hypothesis	• Students who reach an impasse during invention perform better during assessment.	X
	• Reaching an impasse has the largest effect on knowledge that directly resolves the impasse.	X

- **H1: SRL Hypothesis:**

This hypothesis suggests that students become better inventors, and thus are more capable of inventing during the test. One way in which such better domain-general invention skills would be expected to manifest themselves is simply by a higher frequency of actual attempts at inventing on the new-strategy items (as opposed to leaving these items blank). However, analysis of the tests in study 1 shows that advanced students in both conditions attempted 72% of the new-strategy items. This, together with the short intervention, and no explicit emphasize on weak methods, make the SRL hypothesis less likely.

- **H2: Domain knowledge hypothesis**

The Domain knowledge hypothesis suggests that knowledge acquired during invention helps students learn better from the show-and-practice instruction, and that this domain knowledge transfers to the new-strategy items. This hypothesis suggests, among other things, that IPL students are likely to be better at identifying the functional role of the various components of the procedure they were taught (e.g., the sum of distances is divided by N to control for sample size). Specifically, one expects that these students will be better at debugging faulty solution procedures that fail on specific components of the procedure. Indeed, the study found an effect of IPL on advanced students' ability to debug methods (albeit only a marginally significant effect at the *p*=.056 level).

Another prediction that is supported by the domain knowledge hypothesis is that direct mapping between inventions and new-strategy items can be found. In other words, analysis of students’ inventions should demonstrate how key features of the domain can be learned even from faulty inventions. Indeed, an analysis of students' inventions suggests a possible mapping between features encountered during invention and features required for the new-strategy items. Figure 11 shows examples of inventions that may not be mathematically valid, but include a key feature that is required by one new-strategy item. These examples reflect an understanding of the notion that multiple subsets of data can be represented on the same graph. Students who invented these methods may have understood one feature of representations, that is, the ability to compare multiple data sets using a single graph. This idea was not discussed in class, though one of the new-strategy items required its application - it included a histogram with several subsets of data. It may be that experiencing this property of graphs during invention prepared students to better encode it during future assessment.

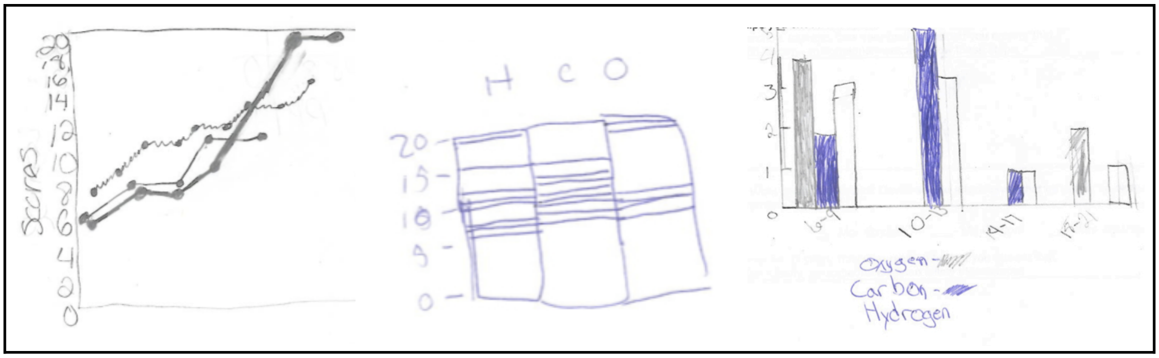


Figure 11: Examples of inventions that include features necessary for new-strategy items. These are three examples for inventions that were used to compare grades of three teachers. On the left: overlapping line graphs. In the middle: each column represents one teacher, each line represent one student. On the right: Four sets of bar graphs in random bins. In common to all these inventions is that they represent multiple subsets of data on a single graph.

The Domain knowledge hypothesis argues that invention modifies the way knowledge is acquired during show-and-practice. Therefore, one expects to find differences also on isomorphic and near transfer measures. However, this was not the case in study 1. According to the study, the effect of IPL is orthogonal to the type of knowledge assessed by isomorphic measures. The significant effect of IPL on new-strategy items with no resources holds even when controlling for performance on isomorphic items on the same test. This shows that the effect of IPL was not merely to scale up learning per-se, but rather, IPL affected the type of knowledge students acquired. Isomorphic measures may not evaluate that knowledge. Students in both conditions learned the material taught during show-and-practice equally well (and this is what is assessed by isomorphic measures). The added value of invention is in students' ability to extend this knowledge to novel challenges.

The improved performance on debugging items and the mapping of features of inventions to features of new-strategy items make this hypothesis likely.

- **H3: Motivational hypothesis**

Existing literature suggests that IPL facilitates several motivational processes that aid learning in general, and far transfer in particular. Specifically, mastery goals may help achieve that effect.

Study 1 found more voluntary work on invention tasks (an indication of greater effort). Given that invention tasks were not graded, this suggests adoption of mastery-orientation goals (since there was no performance-incentive to invest more work on these items).[7] If anything, performance-approach goals would have encouraged students, contrary to what we observed, to invest more time on conventional practice, where more problems could be solved, and more correct application of procedures could be demonstrated. In addition, IPL students reported that the classes during the study were more challenging and demanding, and yet reported to have enjoyed them more. This preference for challenging tasks is an additional sign of adoption of mastery goals.

Alongside the seen motivational benefits of IPL, no correlation was found between increased motivation and performance on new-strategy items. This suggests that while the motivational benefits are important in and of themselves, they are not responsible for the increase in students' knowledge flexibility.

- **H4: Impasse hypothesis**

The Impasse hypothesis suggests that students in IPL realize the limitations of their prior knowledge, and thus are more open to successful learning during classroom instruction. Study 1 provides evidence against this hypothesis. During the intuitive ranking phase, students in both conditions attempted and failed to apply their prior knowledge. (For example, in the topic of variability, students were instructed to attempt average. However, average failed to discriminate between the cases, as became clear in the subsequent class discussion). Given that students in both conditions reached an impasse when applying their prior knowledge, one would expect no advantage for invention.

One may argue that Full Invention students reach a more profound impasse that better prepares them for class instruction. The repeated invention attempts may give students a better understanding of the limitations of their existing knowledge, compared with a single attempt, as

[7] The extra time devoted by the Full Invention students is negligible compared with the overall study duration, and cannot explain the effect of invention in terms of time on task.

was done in the Ranking Only condition. However, in this case, the benefit of Full Invention should be the greatest on isomorphic measures, since these evaluate the knowledge that directly addresses the impasses students reach during invention activities. However, the results of study 1 draw an opposite picture. The fact that students in both conditions had opportunities to reach an impasse, and that hypothetical differences in impasse predict a reverse effect, make this hypothesis not likely.

Q4: generalizability:

The results of study 1 are qualitatively similar to the results published earlier by Schwartz and Martin (2004). Furthermore, these results were obtained in a tighter controlled study, in which the control condition (Ranking Only) shared several key features with the Full Invention condition (i.e., the use of contrasting cases, and the phases of intuitive ranking and short class discussion). This suggests that IPL can be systematically replicated with different populations and researchers, thus confirming our hypothesis.

(b) Bonus track - On future learning measures

The effect of IPL as documented here and in the previous studies is tightly coupled with the new-strategy assessment. It is this assessment that allowed us (and others, e.g., Schwartz and Martin (2004) to identify the cognitive benefits of IPL. The importance of future learning measures can be also seen in the different performance of the two levels of students in the study. Even though students in both classes had previously learned the central tendency materials, the regular and advanced students demonstrated similar performance at pre-test. However, the difference between conditions grew as instruction proceeded (see Figure 6). This, too, suggests that the good students can be characterized not only by their fix knowledge at a certain point in time, but rather, but their ability to learn from instruction.

Chapter 4

Technology for IPL

Facilitating IPL using a tutoring system has several potential advantages for pursuing both engineering and scientific goals. A tutoring system can help us achieve engineering goals by supporting students adequately (and thus reducing the demands form teachers, achieving sustainable improvements in student learning, and supporting scalability). Technology can help us achieve scientific goals by being a research platform on which to manipulate and record the complex interactions that occur during the learning and instruction.

Since facilitating an invention activity differs significantly from facilitating conventional problem solving, teachers would need to undergo extensive training to implement IPL in the classroom. Teachers need to understand the structure of scientific inquiry, desired support during the process, become experts in the domain, and, perhaps most importantly, learn to integrate all of this knowledge in fluent form so they can act quickly enough in real time in the classroom. In addition to workshop sessions, such knowledge can likely only be acquired through practice, like the semester-long effort I engaged in Study 0. Research shows that inquiry tasks often loose their advantages when implemented by insufficiently prepared teachers (Hiebert et al., 2005; Kirschner et al., 2006). Using technology, from a scientific perspective, can help us better understand whether and, more importantly, when and how IPL works so as to know whether and how best to make the investment of teaching teachers this new and complex process. From a practical and policy perspective, technology could eventually help scale these practices by both modeling them for teachers and by easing the demands on teachers so more can successfully adopt IPL instruction.

In addition, an intelligent tutoring system (ITS) for invention can also aid the students. Studies show that ITSs for coached problem solving help students learn better than traditional curricula (Koedinger et al., 1997; Leelawong & Biswas, 2008; Lesgold & Others, 1988; Morgan & Ritter, 2002; Shute & Glaser, 1990). Likewise, an ITS for IPL may achieve greater benefits than paper-based IPL. These benefits can have several forms: they may amplify existing advantages; they can lead to additional (motivational or cognitive) benefits; and they can achieve a similar effect in less time.

Last, an ITS for invention has potential benefits for understanding learning from IPL (and other constructivist) tasks. An ITS for invention can collect detailed log files with detailed information about the interactions students have with the system, allowing for a more fine-grain analysis of learning during invention. An ITS also enables greater control over studies, thus allowing for tighter controlled interventions and within-class manipulations. Last, during the design of ITS, detailed specifications of IPL must be defined, which improves our understanding of the process.

4.1 The challenge: make it intelligent

Much like the IPL process itself, the IPL environment bridges two opposing schools in educational technologies: inquiry environments and ITS.

Inquiry environments support, and at times scaffold, learning from scientific inquiry tasks (de Jong & van Joolingen, 1998; White, Shimoda & Frederiksen, 1999). In these systems students need to uncover an underlying scientific or mathematical model, and in general, behave like scientists do (Scardamalia & Bereiter, 1994). Invention tasks, which share many features with scientific inquiry tasks, could potentially benefit from similar systems. However, classroom evaluations show that students working with inquiry systems often pursue unproductive learning trajectories. Lacking appropriate support, inquiry systems often fail to demonstrate sufficient learning gains (van Joolingen, 1999; Veermans et al., 2000). Overall, while tasks supported by inquiry environments resemble invention tasks, the lack of adaptive support makes these systems less than adequate for facilitating IPL.

ITS, a type of tutoring systems for coached problem solving, seem to offer the right type of support for the tasks they facilitate. ITS use a model of the learner to adapt their feedback and problem selection to the students' actions and knowledge (Corbett & Anderson, 1995). ITS have been shown to increase learning at the domain level (Koedinger et al., 1997; VanLehn, Lynch, Schulze, Shapiro & Shelby, 2005). Yet, there is considerable evidence that students demonstrate poor metacognitive behavior while working with ITS (Aleven & Koedinger, 2000; Koedinger, Aleven, Roll & Baker, in press). Though research suggests that students can benefit from monitoring their own learning behavior within ITS (Aleven & Koedinger, 2002; Mathan & Koedinger, 2005; Reif & Scott, 1999), they are rarely required to do so. As with direct instruction, students often do not get to practice their inductive and scientific reasoning skills.

ITS has many features that make it suitable for invention tasks. For example, the cognitive model in the basis of ITS can evaluate classes of equivalent solutions, rather than individual solutions (for example, a single rule can evaluate equivalent fractions such as 1/3, 2/6, 50/150, etc). However, since invention tasks are more open ended than traditional ITS tasks, some difficulties arise when adapting ITS to IPL. In order to give meaningful feedback, the ITS should be able to interpret and evaluate students' solutions. Both these tasks are not obvious in the case of a scientific inquiry environment, as explained below.

A main source of difficulty is the infinite interaction space in invention and other inquiry tasks. By interaction space I refer to the collection of potential actions that students can perform and the system needs to parse and interpret. ITS often constrain the interaction space. For example, some tutors include menu-based selections (Aleven & Koedinger, 2002). This limits the number and variability of inputs that the system should interpret. Also, most ITS give students immediate feedback on their errors, which helps students stay on course. Furthermore, most systems require students to fix their errors as soon as these are identified. Interface structuring, immediate feedback, and constraints on allowable moves can be used to better predict, at each moment, a more limited (and thus more computational feasible) set of legitimate subsequent actions. These characteristics of the interaction with ITS allow tutoring systems to provide practical solutions to the otherwise NP-hard problem of plan recognition. In inquiry systems, on the other hand, students are typically much more free to roam the interaction space, including going down unproductive paths. In the variability task, for example, every mathematical method is a legitimate response. Furthermore, the system should not evaluate the numeric answer generated by the methods, but instead it should evaluate the method itself. This means that not only there are many more possible inputs, but they also vary a lot in their structure and complexity. The problem of creating a cognitive model to accurately interpret all possible inputs within the interaction space with respect to a global plan (i.e., perform plan recognition) is nearly impossible in practice.

Another source of difficulty in using ITS to support inquiry tasks relies on the under-defined solution space of invention and other inquiry tasks. By solution space I refer to the variety of responses that the system should evaluate (i.e., assign truth value to). Earlier I discussed the difficulty in interpreting the method. However, even once parsed, the correctness of the answer should be evaluated by the tutoring system, and in case of some partial or incorrect responses, the system should also identify the source of error. To do that, ITS use a set of rules that evaluates all common solution paths (the cognitive model). These rules include typical correct responses, as well as other partial and buggy responses that have pedagogical significance. Furthermore, most ITS make an assumption that a solution that is not captured by a correct rule is incorrect. However, the variability in responses in invention tasks makes it hard to characterize classes of solutions in advance. For example, applying range as a measure of variability is a common misconception, and thus should be included in the model. One of the students in study 1 invented "range + 1". A conventional cognitive model would not have identified this to be a variant on range, since it is impractical to include all variants on range in the model. Furthermore, invented methods may share features with range, but behave differently when applied to different contrasting cases – such as another method that was invented in study 1, "min+max". Creating a model that is flexible enough to identify these methods, without predefining them, is not easy to create, has rarely been implemented, and even more rarely implemented in classroom settings.

4.2 Existing systems

Several systems have tried to incorporate intelligent feedback into inquiry environments. Among the more influential ones are SmithTown (a simulated economic market, aimed at teaching microeconomics and scientific inquiry skills; Shute & Glaser, 1990); Rashi (a domain-independent scientific learning environment, with tutoring modules in the fields of health, science, and engineering; Woolf et al., 2003); SimQuest (a discovery environment for physics; Veermans et al., 2000); EcoLab (a simulated ecology laboratory for elementary school students; Luckin & du Boulay, 1999); Science Learning Spaces (a simulation environment with intelligent feedback for science teaching; Koedinger, Suthers & Forbus, 1999); and Crystal Island (a dialogue-based inquiry game for middle school biology; Mott & Lester, 2006). Most of these systems were evaluated in controlled studies; all those that were evaluated succeeded in helping students learn. An analysis of these systems revels several commonalities in the ways they support the inquiry process.

In order to deal with the complex interaction space problem, and to be able to parse the different inputs, most systems limit the interaction space in one way or another. One common way of doing so is to scaffold the scientific inquiry process itself. For example, Rashi, Smithtown, and the Science Learning Spaces include an inquiry notebook with templates in which students are prompt to raise hypotheses, document observations, make conjectures, etc. (Koedinger et al., 1999; Shute & Glaser, 1990; Woolf et al., 2003). Similarly, SimQuest has tools for collecting data and raising hypotheses (Veermans et al., 2000). Scaffolding the inquiry process helps tracking students' progress on the task. In addition, such scaffold gives students additional instructional assistance. For example, by making thinking visible, students are more likely to internalize these processes (Collins, Brown & Holum, 1991; Roll, Aleven, McLaren & Koedinger, 2007).

Another mechanism by which systems limit the interaction space and make interpreting of solutions easier is by narrowing the vocabulary students can use, either by using menus or predefined variables. For example, several systems ask students to state their hypotheses using built-in variables (Shute & Glaser, 1990; Veermans et al., 2000). Other systems control the data that is available to students. Rashi allows students to read scientific papers, but only papers that are mapped onto the system's database (Woolf et al., 2003). Similarly, Crystal Island lets students interview experts, but only within the game (Mott & Lester, 2006). Doing so reduces the number and variability of responses that the system needs to interpret. In addition to constraining the interaction space, these scaffolds offer additional instructional assistance that may improve learning.

Though inquiry systems usually do not give immediate feedback at the domain level, two other forms of feedback or commonly used. One is process level feedback. Inquiry systems can use their scaffold to evaluate the inquiry activities themselves in a domain independent manner. For example, Rashi gives feedback to students who make circular arguments (Woolf et al., 2003), and Science Learning Spaces gives feedback on experimental designs that do not use the control of variables strategy (Koedinger et al., 1999). In addition, many inquiry systems include situational feedback (Nathan et al., 1992), in which the environment gives students implicit feedback (by behaving in a manner that deviates from students' expectations). A similar technique was also found useful in ITS, by giving students opportunities to diagnose their own errors, and as long as these have clear perceptual characteristics (Mathan & Koedinger, 2005).

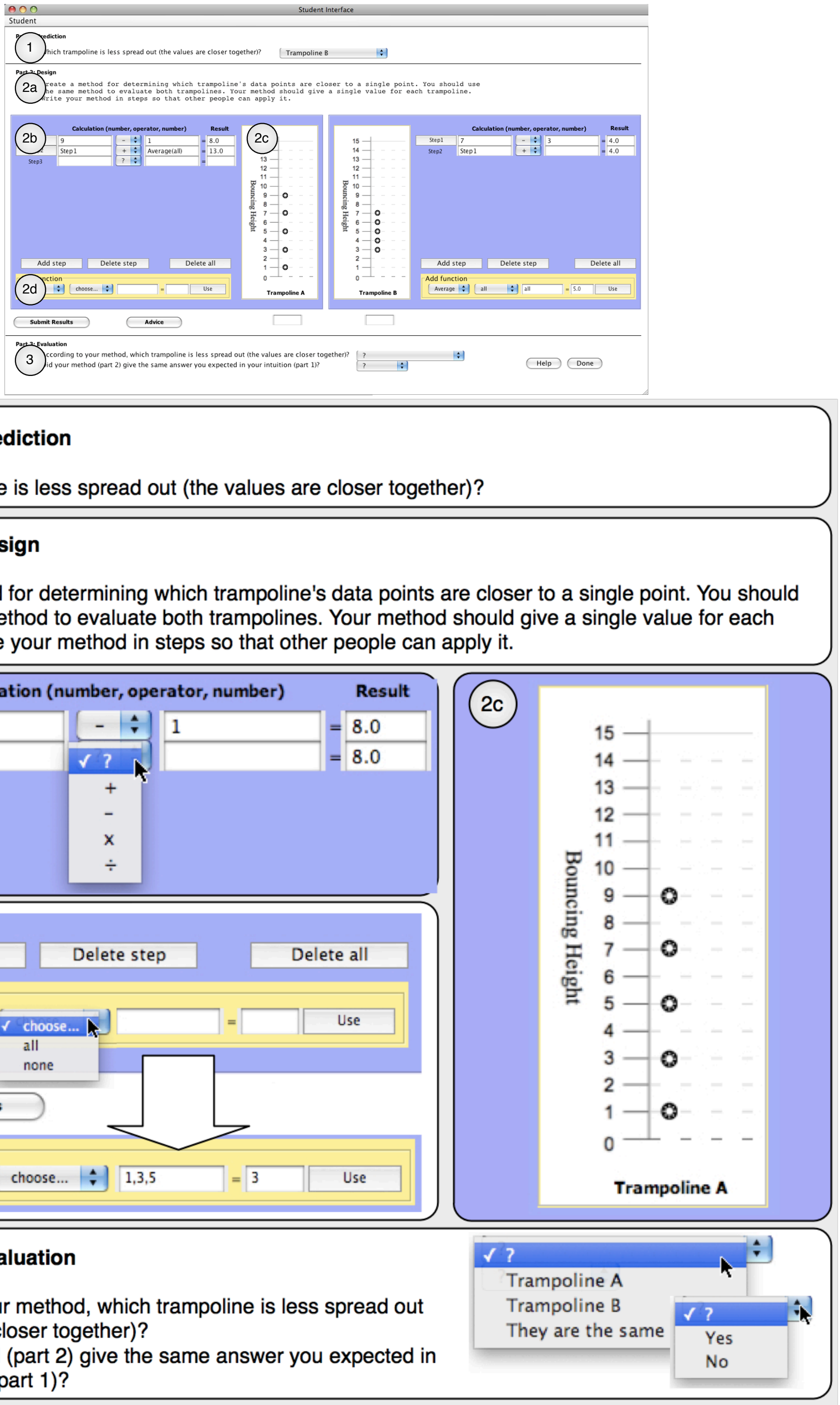
The other challenge intelligent inquiry systems face is the under-defined solution space. Different approaches tackle this challenge in different ways. Several systems put sufficient constraints on the interaction space, and thus are able to predefine all relevant solutions (usually correct solutions) either through example-based enumeration (sometimes easier) or though abstract representations (many-to-one mappings, rules, schemas). For example, SimQuest uses the finite number of potential hypotheses to evaluate the complete subset of potential experiments (Veermans et al., 2000). Similarly, EcoLab maps all possible nodes in the system, essentially having a predefined reaction to each potential state and user level (Luckin & du Boulay, 1999). Other systems simplify this task by having only one specific target solution that students need to reach (Crystal Island, Mott & Lester, 2006, and EcoLab, Luckin & du Boulay, 1999). A different potential solution does not evaluate the solutions themselves, but rather, evaluates

4.3 The Invention Lab

The Invention Lab is an ITS for IPL, built using the Cognitive Tutor Authoring Tools (Aleven et al., 2006). Given that the invention process resembles a scientific inquiry, the system can be viewed as an intelligent inquiry environment. In it, students inquire about the phenomena of data spread or variability. Notice, though, that the invention process does not include all components of scientific inquiry. For example, the system does not ask students to collect data. Instead, it presents them with carefully designed and systematically chosen contrasting cases, and it does the calculations for them. The Invention Lab can evaluate any mathematical method that students construct using its interface, without having to pre-define the solution space. However, its uniqueness is not in the methods it finds valid, but rather, in the unsatisfactory methods. By analyzing students' faulty inventions the Invention Lab evaluates students' understanding of key features of the domain (in this case, variability). Rather than merely identifying errors, the lab can identify specific classes of errors (such as not controlling for sample size), and thus give adaptive support to students.

The Invention Lab achieves that using two cognitive models: A meta-model of the task, and a domain-level model of variability. The lab was designed to be used by pairs of students sharing a computer, but can also be used by individual students.

The Invention Lab facilitates the invention activity that was used in study 1 (intuitive ranking, design, evaluation). Figure 12 shows the Trampoline problem in the Invention Lab. This problem (which was used in studies 1 and 2) asks students to invent a method for identifying which of two given trampolines is more fair to all athletes (see Figure 1). The two graphs in the center show the two contrasting cases, that is, a data pertaining to each trampoline. In this problem the contrasting cases present data from bounciness tests. Each graph shows how high a standard weight bounced when dropped on the trampoline in identical conditions. Each point represents the first bounce in a different trial (i.e., Trampoline A was tested five times. The first bounces in these five tests were 1, 3, 5, 7, and 9 units high).



Part 1: Prediction

Which trampoline is less spread out (the values are closer together)?

Part 2: Design

Create a method for determining which trampoline's data points are closer to a single point. You should use the same method to evaluate both trampolines. Your method should give a single value for each trampoline. Write your method in steps so that other people can apply it.

Calculation (number, operator, number)				Result
9	−	1	=	8.0
Step1	✓ ?		=	8.0

Add function: Count · choose... · 1,3,5 = 3 · Use

Part 3: Evaluation

According to your method, which trampoline is less spread out (the values are closer together)?
Did your method (part 2) give the same answer you expected in your prediction (part 1)?

Figure 12: The Invention Lab interface. On top - the lab; on the bottom - its components. Students begin the invention activity by ranking the contrasting cases (1). Upon successful ranking, the system asks the students to invent a method that will reflect their ranking (2a). The students express their method in steps (2b), using points from the contrasting cases (2c), and using basic functions (2d). Last, students evaluate their method (3) and revise it as needed. The lab is designed to be used by pairs, though this is not a constraint.

Students begin their invention activity by reporting the outcome of the intuitive ranking, that is, choosing which case is better according to the predefined criteria (in this task, which trampoline is more fair; region (a) in Figure 12). The lab gives immediate feedback on students' observations. Ranking the cases correctly is important for two reasons. First, if the intuitive ranking is wrong, a teacher intervention is needed to explain the concept. Second, the intuitive ranking serves as the baseline for evaluation of the method.

Upon successful intuitive ranking students move on to the design phase (region (2a) in Figure 12). In this phase, students design a method to calculate the consistency of the trampolines (as a quantitative measure of their fairness). To support students' natural tendency, designing such a method in the lab is done in steps (2b). Each step has the simple form of number -> operator -> number. Students can click on points on the graph to enter their values (2c). For example, in order to get "9-1", students click on 9, choose the minus sign, and click on 1. Students can also use basic functions (sum, average, median, and count) from the yellow panel at the bottom (2d). Last, students can use the results from a previous step in a current step. For example, Step 2 in the magnified method uses Step 1. To reduce the cognitive demands during invention, calculations are carried out by the system.

Notice that while students need to invent a general method, they need not express it as such. Instead, they instantiate it right away to the given cases (c.f., Heffernan & Koedinger, 1997; Koedinger, 2002) This was done based on the lessons learned from studies 0 and 1 and to reduce the need for using complex symbols (such as sigma or parenthesis). For example, inventing range is done by clicking on the top point in the relevant graph, choosing the minus sign, and then clicking on the bottom point in the same graph. Students work simultaneously on applying their method to both cases. Once done, students submit their method, which ends the design phase.

No immediate feedback is given during the design phase. However, the Invention Lab checks for consistency upon submission of the method, that is, the its checks whether the same method was applied to both contrasting cases. This constrain on students' solutions is commonly used by constraint-based tutors, which, instead of specifying correct solutions, specify list of constraints that should be satisfied by the solutions (Mitrovic & Ohlsson, 1999). When different methods are applied to both contrasting cases, the lab points it out to the student. When the method is consistent, the student moves on to the evaluation phase.

The evaluation phase asks students to compare the outcomes of their methods to their predictions (region (3) in Figure 12). Feedback on evaluation incorporates an intelligent novice model, in that students are first given the opportunity to notice the limitations of their methods and revise them (Mathan & Koedinger, 2005). Failing to do so triggers feedback from the lab. When the method generates the right prediction for the given contrasting cases, the system identifies its missing features and generates appropriate contrasting cases (this process is further detailed in the next section). When the model identifies no missing features, a generic set of contrasting cases (chosen randomly from a bank) is presented to the student. Following that, students begin a new cycle by ranking the new set of contrasting cases.

Each exercise in the Invention Lab includes several cycles of intuitive ranking -> design -> evaluation. The tasks are designed to engage students for 30 minutes, though no time limit is enforced by the system. Currently there is no exit-point from the cycle, since students are not expected to design valid methods and rarely succeed in doing so (as seen in studies 0 and 1 and reported by Schwartz and Martin, 2004).

Since the invention lab interface is unique, a short tutorial was developed and user-tested. The tutorial instructs students how to find the average of data in three different ways (the long way - adding all points and dividing by N; a middle way - using the 'sum' function and dividing by N; and the short way - using the Average function in the lab). This process includes all major components of the lab. The tutorial uses the same cover story as the first invention task - the trampolines exercise. By instructing students to use average, the tutorial helps them notice that average does not suffice, without priming specific options for variability.

4.4 Achieving intelligent interactive instruction in the invention Lab

To give intelligent feedback, the Invention Lab uses several mechanisms. First, it scaffolds the inquiry process by splitting the task into its three phases: intuitive ranking, design, and assessment, and clearly indicating those phases in the interface (see Figure 12). The lab also traces students' progress using a cognitive model (written in Jess, Friedman-Hill, 2003), and gives feedback on domain independent inquiry errors. For example, when students fail to notice that predictions derived from their design do not match their observations, the tutor responds by explicitly pointing out that “but your answer in the last question is not the same as your initial prediction." The lab also gives feedback on general mathematical errors, such as inconclusive methods. Students' methods should give a single value for each graph, and failing to do so triggers appropriate feedback. For example, when students do not connect the different steps included in their method, the lab responds by instructing them to “…involve each of your steps in the final step you use. In the left graph, step 1 is not used to compute the final result."

Second, as other inquiry systems have done the lab limits the interaction space by privileging some forms of desirable action but not allowing others that might be used on paper or outside the computer environment. In particular, use of step-by-step arithmetic expressions to express inventions is privileged, but students cannot use free text or diagrams within the invention lab interface. Of course, they can in principle still use paper to write or draw and they do engage in partner discussions, but they do not get feedback on these actions (nor are they automatically stored in the record). Besides making students’ methods more interpretable, limiting students to mathematical notations was shown to have positive effect on learning (Schwartz et al., 2005). In an attempt to minimize the limitations on the interaction space, the lab allows for a variety of mathematical expressions (at the middle school level). For example, step 1 in Figure 12 includes the range function. The lab traces the exact actions and stores the method in its working memory. In this example the system knows that the student subtracted the lowest value from the highest one. This information is used to build new contrasting cases, as explained below.

The main intelligent component of the Invention Lab is its representation of domain knowledge, which allows it to analyze novel solutions and trace students' knowledge levels. While the mechanisms detailed above assist the lab in interpreting students' method, they are not sufficient to evaluate the correctness of the methods or analyze their features. Also, none of the approaches described above for solving the solution space problem supports analysis of all possible inventions. The numeric answers given by the methods are not informative enough, and, given that students' methods are often under-defined and cannot be generalized, a cognitive model cannot evaluate whether a method is globally correct (i.e., for all possible cases).

Unlike conventional cognitive models that evaluate the correctness of the students' solution steps and final entries (Corbett & Anderson, 1995), the cognitive model of the Invention Lab evaluates the deep features of the invented methods, much like the constraints in Mitrovic and Ohlsson (1999). Extending the intelligent novice model approach (Mathan & Koedinger, 2005), the lab only evaluates local correctness of the method (i.e., correctly ordering of the two current cases) at a delay, that is, after the student has explicitly performed a self-evaluation of local correctness. To do that, the knowledge base of the Invention Lab does not include procedures, but it includes features of procedures. Although the number and diversity of possible methods is unlimited, the number of key conceptual features is finite, making it tractable. For example, many methods that students create use only the extreme data points to determine spread (e.g. range, or adding the distances from top to average and from bottom to average). Yet, all these methods reveal the same incorrect notion that variability can be determined only by the data's extreme values. The Invention Lab need not represent all the possible ways of using only the extreme values (and students generate many). Rather, it can simply identify when the only arguments used by a method are the extreme values. A comprehensive list of 6 target features with 14 associated errors was compiled based on students' inventions during study 1. Table 14 shows a subset of these features, and appendix 1 includes the comprehensive list.

Table 14: Identifying conceptual errors in procedural methods. This is a simplified subset of the full cognitive model, which includes 6 target components with 14 common conceptual errors.

Target feature	Common associated conceptual error	Examples of methods invented by students, applied to sample data (2,4,4,7,8)				
		Range times two	The number of different values	Average divided by number of data points	% of points close together	Largest gap between subsequent points
		(8-2)*2 = 16	Count (2,4,7,8) = 4	Average(all) / count(all) = 1	Count (2,4,4) / count(all) = .6	Ga*p*(4,7) = count(5,6) = 2
Variability is determined by all points in the data	• Method uses only extreme values	X				
	• Method uses only a sequential subset of points.				X	X
Variability is not central tendency	• Method uses only measures of central tendency.			X		
The method should control for sample size	• Method does not control for sample size.	X	X			X
Repeated values should be taken into an account	• Method uses only the gaps					X
	• Method ignores repeated values		X			
Variability depends on distances	• Method does not use distances between points		X	X	X	
Only given data should be used	• Method uses arbitrary constants	x				

To the extent that every mathematical method (at the middle school level) can be expressed using the Invention Lab interface and every invention can be analyzed according to the features described above, the cognitive model of the lab can give intelligent feedback on any method, thus, de facto, not constraining the solution space. The features used by the invention lab are parallel to facets, as used by Minstrell (2001), in that students' answers are used to identify lacking features, which, in turn, trigger additional problems targeted at the demonstrated conceptual knowledge gaps. The features of the Invention Lab also resemble constraints as used by Mitrovic and Ohlsson (1999), in that the lab evaluates whether the solution adheres to several governing rules, rather than evaluating the solutions themselves. We use conceptual feature [if you want to stick with that] because it corresponds with result that a critical consistent difference between experts and novices is their ability to extract the deep solution-relevant features from problems and scenarios (Chi, Feltovich & Glaser, 1981).

Like other ITS, the Invention Lab uses its evaluation of the students' knowledge to choose the subsequent challenge (Corbett & Anderson, 1995). However, choosing contrasting cases that will challenge students' methods is no easy task. Since students instantiate their methods using the contrasting cases, and do not express a general rule, it is hard to generalize their method and predict their outcomes on new contrasting cases. Students' methods tend to be under-defined, especially that they may not include features that are not emphasized by the contrasting cases (for example, generalizing from two cases with equal N to new cases with different N can be done in more than one way). Instead, unlike most other ITS, rather than choosing a pre-designed set of contrasting cases, the Invention Lab designs in real time a new set of contrasting cases to match the student's needs. First, the Invention Lab chooses a domain feature that the student failed to show proficiency on (i.e., their method did not take into account this feature). When there are multiple candidates, the lab selects a feature based on a predefined prioritized list. The system targets each common error up to three times. If students persist in violating a certain feature after five sets of cases the system moves on to the next feature. Then, the system generates a set of contrasting cases. Each common error has an associated method for generating new contrasting cases (see Appendix 1). The process is designed to ensure that new sets of cases are easy to compare with regard to target concept (so intuitive ranking will be simple), and that the most recent method would fail on them. For example, if the student used only extreme values (e.g., range), the system will generate two new cases that share the same range but have different variability. Last, the process uses the recent set of cases, to help students build upon their prior experiences and methods, and thus create more cohesive knowledge (rather than a collection of ad-hoc methods). Table 15 demonstrates this process, which is detailed in Appendix 1. First, the system keeps the values that the student used from one of the cases in the previous set, often keeping one of the cases intact. The values that are kept from the previous contrasting cases are called *seed*. Keeping the seed is done to help students transfer their method from the previous set to the current one. The seed is often the case with the higher variability (or the "loser") from the previous set of cases, to leave more room for manipulations (since the "winner" case often has points that are too close together to create interesting contrasts). The second step is to make sure that the previous method fails to distinguish between the current contrasting cases. This is usually done by using the same seed for both contrasting cases. If, for example, the previous method used only extreme values (e.g., range), using the same seed makes sure that the extreme values of both cases in the current set are identical. While the goal of step 2 is to make sure that the cases are identical with regard to features that were exploited by the student, the goal of step 3 is the opposite - to make sure the sets differ with regard to the new target feature. In this step the system 'populates' the cases with up to 6 points, in order to focus students' attention on the new target feature. There are different ways to populate the cases, based on the target feature. For example, if the target feature is controlling for sample size, step 3 will make sure the cases have different number of points while keeping the same average and range. It usually achieves that by using each value in the seed twice (so one case is exactly twice the other case)

Once the system is done generating a new set of cases, it calculates the MAD for both cases. The difference between the MADs of the contrasting cases should be greater than 1.2 units. This value was chosen by testing the minimum difference between MADs that is still perceptually noticeable. It is important insure that students can rank the cases correctly during the intuitive ranking phase. If the MADs are not far enough, the system generates new contrasting cases, and check this exit condition again. After 500 attempts the system does not attempt more cases, and instead picks a random set from a pre-defined collection of sets for each target feature.

Table 15: Simplified example of real-time generation of contrasting cases. The process creates cases that are indistinguishable by the previous method used by the students, while directing students' attention to additional features of the domain.

	Case A	Case B	Comments
Original task			
Original cases:	2 3 4 7 9	3 4 5 5 6	
Original invention by student: range	9-2 = 7	6-3 = 3	Student uses range, and determines that set B has a lower variance
New task			The lab chooses to focus on the following feature: variability is a function of all points in the data.
1. Keep the case with the higher variability from the previous cycle	2 3 4 7 9		This encourages students to transfer from previous experiences
2. Include the values that were used by the student in her previous method (i.e., seed) in both cases	2 3 4 7 9	2 9	This ensures that the pervious method fails to distinguish between the cases in the new set.
3. Populate case B with values that are halfway between the average and the values of case A.	2 3 4 7 9	2 4 5 6 9	This ensures that the two sets have distinct variability, easy to judge perceptually. (original average = 5. halfway between 3 and average is 4 halfway between 4 and average is 5 halfway between 7 and average is 6)

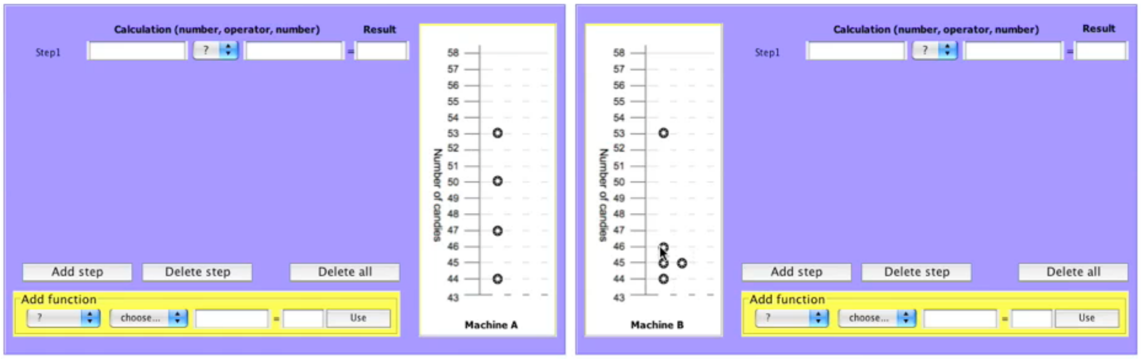
4.5 Initial evaluation

The Invention Lab and its tutorial were developed with frequent and iterative user testing. A close-to-final version was tested with 7 middle- and high-school students. Think aloud protocols were collected, though due to poor audio quality not all conversations are comprehensible. Students completed the lab tutorial (identical to the one that was later used in study 2) and then used the lab for 2-3 hours (excluding other instructional time). The Invention Lab seemed to do at least as good as a paper-and-pencil version of the invention activity. Students were engaged in the process, and designed inventions that were qualitatively similar to inventions observed during study 1 (and some of them were literally identical). Students appeared comfortable with the novel interface, and I observed only few difficulties entering the methods they expressed verbally. Cases generated by the lab were successful at emphasizing the target features, as was evident from students' comments, and students identified their conceptual errors as they applied their previous methods to new cases generated by the system (e.g., they found that the method did not produce the intuitive results predicted during the intuitive ranking phase). Last, students' success rate during intuitive ranking was high.

Along side the positive findings, the user testing revealed some difficulties that were later seen also in Study 2. Even though the students were able to use the interface, it was too time consuming to enter invented methods this way, and many students opted to use paper (or mental math) to initially generate a method. Students often used the paper to evaluate their methods prior to entering them into the computer. In addition, students were more eager to crack the cases - that is, to find methods that produced a correct ranking for the displayed contrasting cases, than they were focused on creating methods with an appropriate justification (see Table 16). Klahr and Dunbar (1988) describe two spaces that students can explore when trying to construct scientific models. In the Hypothesis space students begin from justifiable hypotheses. In the Experiment space students tend to use a more engineering approach that is closer to trial and error. While Study 1 had many students exploring the hypothesis space, the Invention Lab seemed to encourage students to explore the experiment space.

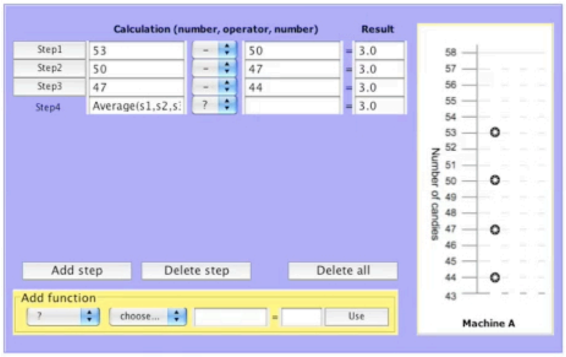
Table 16 shows an example from one of the pilot studies that demonstrates many of the points mentioned above. This example features a boy and a girl working together. Both are very good at math in their middle school, but have never learned about variability. They already worked for 30 minutes on the first invention task (trampolines) and we join them in minute 14 of the second invention task (evaluating the consistency of machines that pack candies). Points in the graph represent how many candies are in each package. The boy is holding the mouse.

Table 16: Annotated snippets from the initial user testing of the Invention Lab.

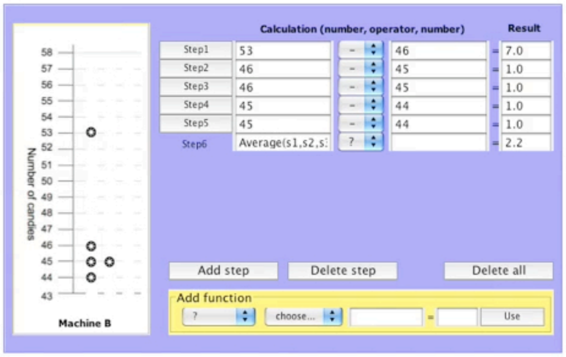


Contrasting cases: (44, 47, 50, 53) vs. (44, 45, 45 ,46, 53). The cases differ on several dimensions since these students have already cleared many misconceptions.

Dialogue	Comments
Boy: *Well, should we try to use this method that we used before?*	Integrating knowledge: Referring to their previous method, average distance between subsequent numbers
Girl: [Calculating the average distance between subsequent points for the left graph using mental math]: *Well, this is just going to give us 3.*	Using mental math before typing anything in the system.
Boy: *3, right*. [begin calculating the right graph by adding up the distances]. *This is going to be... 11. 11 divided by 4. Wait, divided by 5. Oh, we have 5 with the one from there, and the one from there* [pointing to the two distances on each side of the repeated value]. *and it is going to be 11...*	Noticing feature: Boy tries to calculate average distance, but is not sure how many distances there are, since it is not clear how many distances to count from the repeated value.
Girl: *These are 5 points. There are 5 points. So if there are 5 points we should divide by 5.*	
Boy: *No, if there are 5 points, it does not mean 5 spaces. So if you have.... 5 points, 4 spaces.*	Noticing feature: Realizing that the number of distances is not the number of points
Girl: *But in this one [points to the right graph] we got 2,3,... 5 spaces. Right?* Boy: *So maybe that's what is wrong with the method. there...*	
Girl: *Three spaces on the left side* Boy: *Right* Girl: *This one [pointing to the right case] has more points, so it is okay if it has more spaces*	Resolving conflict: Noticing that number of points is not the same in both graphs.
Boy: *But it has only one more point and two more spaces. So there should be a mathematical trick.*	Noticing that the discrepancy between points and distances is not identical, and becomes suspicious of the method.
Girl: *So here if we use the 5 spaces we are going to end up with 2.2 and here it looks like the values are closer together except for the one that is far away. But I like this one better because it is more concentrated.* *[*They enter their method, and it yields the correct ranking for the given cases.*]*	Switching to experiment space: Girl is calculating to convince boy that it works.

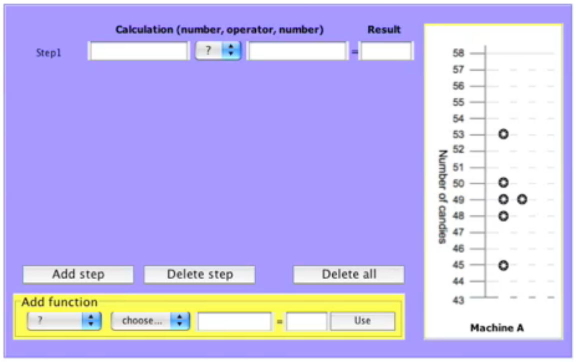


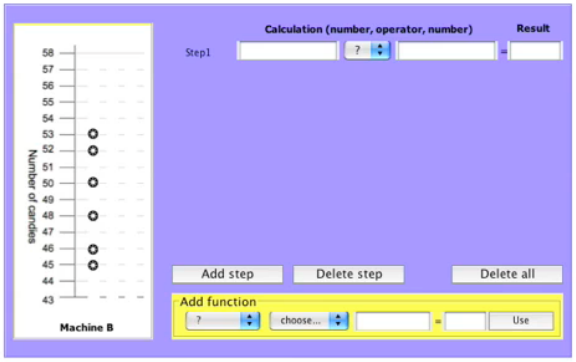
Calculation (number, operator, number)				Result
Step1	53	-	50	= 3.0
Step2	50	-	47	= 3.0
Step3	47	-	44	= 3.0
Step4	Average(s1,s2,s:	?		= 3.0



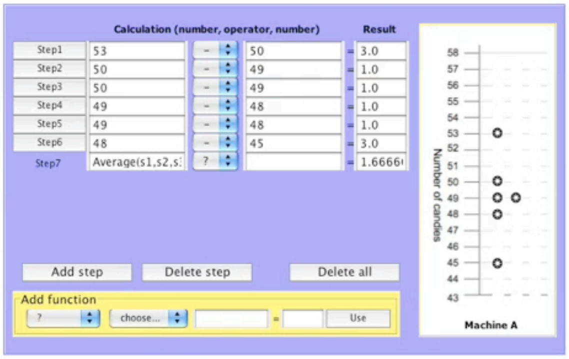
Calculation (number, operator, number)				Result
Step1	53	-	46	= 7.0
Step2	46	-	45	= 1.0
Step3	46	-	45	= 1.0
Step4	45	-	44	= 1.0
Step5	45	-	44	= 1.0
Step6	Average(s1,s2,s:	?		= 2.2

Next the system gives the following cases: (45, 48, 49, 49, 50, 53) vs. (45, 46 48 50 52 53).

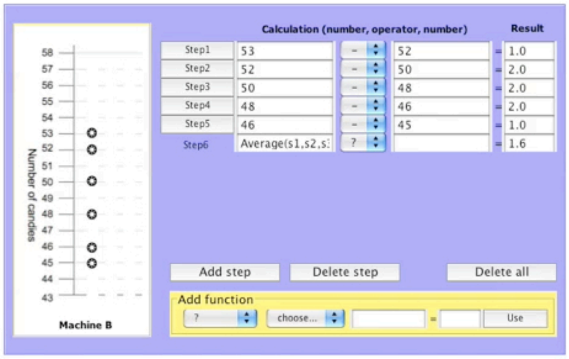




The students try to apply the same method, but this time it fails.



Calculation (number, operator, number)				Result
Step1	53	-	50	= 3.0
Step2	50	-	49	= 1.0
Step3	50	-	49	= 1.0
Step4	49	-	48	= 1.0
Step5	49	-	48	= 1.0
Step6	48	-	45	= 3.0
Step7	Average(s1,s2,s:	?		= 1.6666



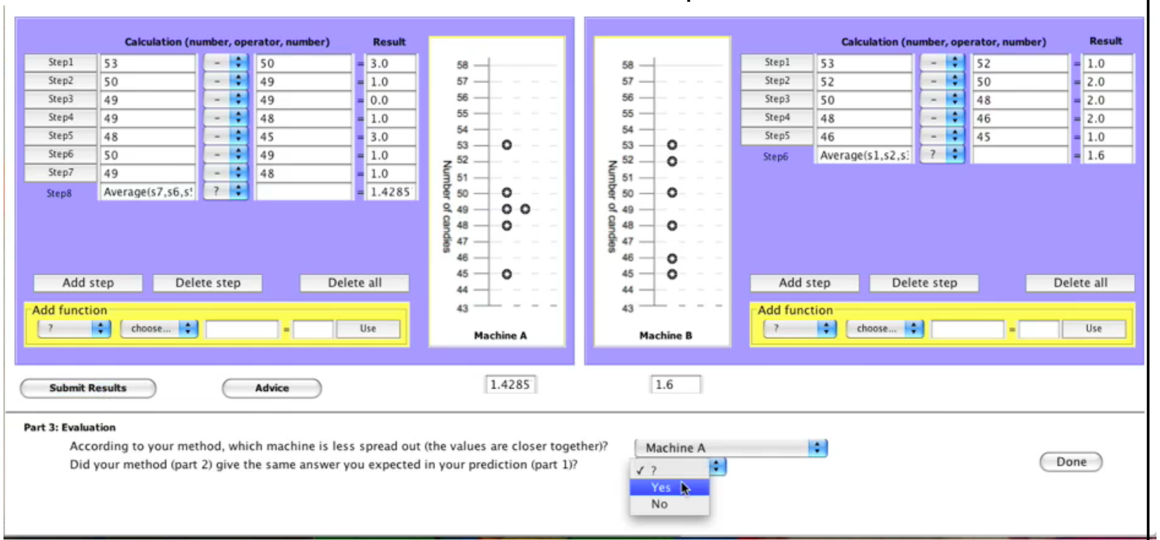
Calculation (number, operator, number)				Result
Step1	53	-	52	= 1.0
Step2	52	-	50	= 2.0
Step3	50	-	48	= 2.0
Step4	48	-	46	= 2.0
Step5	46	-	45	= 1.0
Step6	Average(s1,s2,s:	?		= 1.6

Girl: *Huh, so I guess this is wrong*

Boy: *So should we delete all our stuff and try the first stuff just in case? Ahhh..., alright. So remember - this is just 1.6* Girl: *Do you want to try the first one?*	Failing to understand generalizability: deciding to attempt a previous method that worked at first but later failed.
Boy: *Yeah...*	
Boy: *So we are aiming for a number below 1.6... Our first method was just instead of going here and there just to go there there and there...* [pointing at different points and suggesting to change the method] Girl: *Yes, it is going to be a 0 instead of being two 1's. So now we have to make it so that it brings...*	Experiment space: attempting various calculations in order to get low result. They are no longer concerned about what is right to do.
Girl: *Yes, it is going to be a 0 instead of being two 1's. So now we have to make it so that it brings...* Boy: *Make it lower...* Girl: *This will give us zero* Boy: *We will do everything to...* [students laugh]	Adoption of performance goals: Students attempt to crack the problem, not to invent a method.

Girl: *Let's see if we can tweak that instead of tweaking that* [suggesting to revise intuitive ranking]
Boy: *Unless the candy company wants to make big packages and small packages* [pointing at the graph with the bimodal distribution; students laugh]

Once their method fails, they consider revising their ranking to match their method. At the same time, they realize that their ranking is correct and should not be altered.



	Calculation (number, operator, number)				Result
Step1	53	-	50	=	3.0
Step2	50	-	49	=	1.0
Step3	49	-	49	=	0.0
Step4	49	-	48	=	1.0
Step5	48	-	45	=	3.0
Step6	50	-	49	=	1.0
Step7	49	-	48	=	1.0
Step8	Average(s7,s6,s!	?		=	1.4285

Machine A: 1.4285

	Calculation (number, operator, number)				Result
Step1	53	-	52	=	1.0
Step2	52	-	50	=	2.0
Step3	50	-	48	=	2.0
Step4	48	-	46	=	2.0
Step5	46	-	45	=	1.0
Step6	Average(s1,s2,s:	?		=	1.6

Machine B: 1.6

Part 3: Evaluation

According to your method, which machine is less spread out (the values are closer together)? Machine A

Did your method (part 2) give the same answer you expected in your prediction (part 1)? ✓ ? / Yes / No

Eventually they succeed in tweaking their method so it ranks the contrasting cases appropriately, and continue to the evaluation phase.

Boy: *Are you ready? here we go...* [students submit method, the lab confirms their success.]
Girl: *YES YES!*
Both: [high-five] *ALRIGHT!*

Boy: *That's hard core math. that's some hard core math*
Girl: *That is so cool. That feels so good. That is so satisfying!*
Boy: *That IS so satisfying!*

Sense of ownership and higher self-efficacy.

Chapter 5
Study 2: Evaluating the Invention Lab

5.1 Focus

The goals for Study 2 were to evaluate the general feasibility of the Invention Lab, and thus address Q4 (technology and IPL) and to use it to further investigate what components of IPL are critical. I hypothesized that the Invention Lab would succeed in facilitating IPL, that is, would allow students to construct inventions and improve robust learning from subsequent instruction.

With regard to Q1, identifying the critical components of IPL, study 1 found that some combination of design and evaluation, on top of intuitive ranking, is necessary. Study 2 further examines the roles of design and evaluation. It does so by contrasting a full IPL condition with two alternative system versions that include intuitive ranking and evaluation but not design. Study 2 also addresses Q3, Cognitive Mechanisms, by directly evaluating several of the predictions stated in Table 3. The study specifically evaluates the Domain Knowledge hypothesis, by including a condition in which students are explicitly told the different domain features.

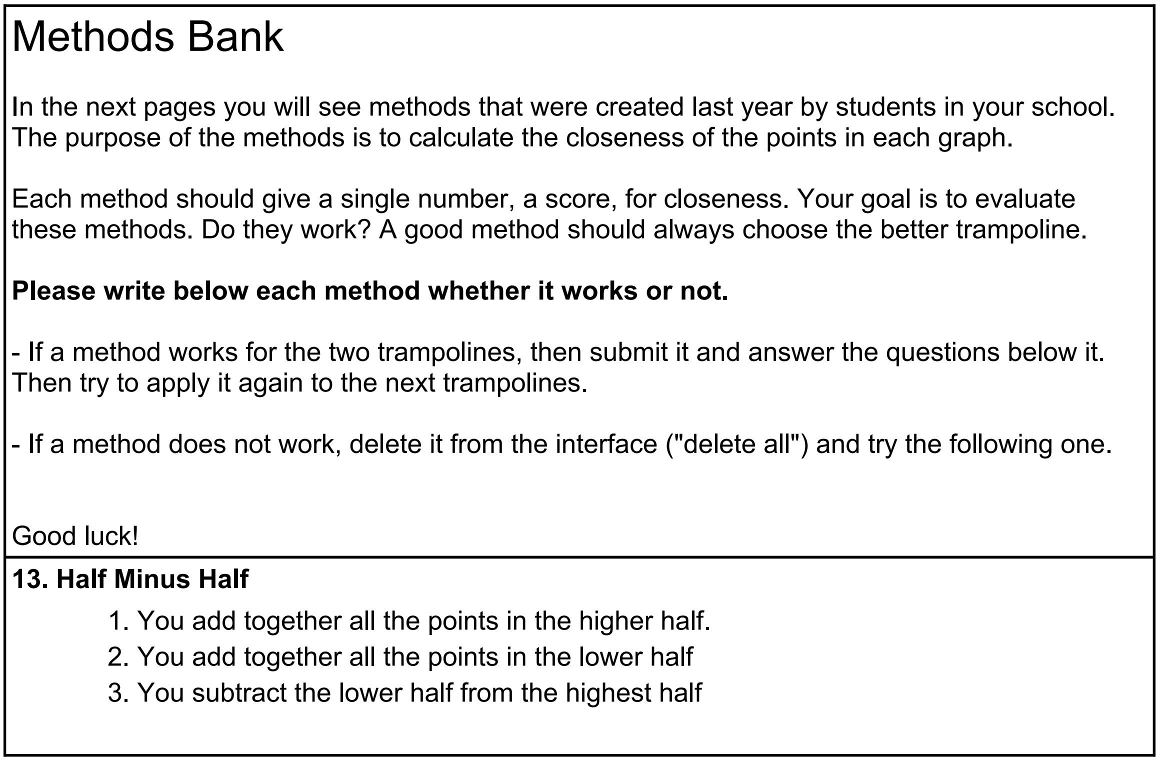
Study 2 also elaborates the answer to Q2, mapping the overall effect of IPL. The study assesses metacognitive behavior during invention within the Invention Lab, as well as invention behavior in an isomorphic problem in a new domain during post-test.

The study includes the following three conditions:

1. Full Invention: Students in this condition were engaged in the full IPL process, similar to Study 1. This is the only condition in which students designed their own methods. Greater learning gains in this condition would suggest that students own design of methods is a critical part of IPL.
2. Method Evaluation: Students in this condition were given pre-designed methods, and were asked to apply and evaluate them. The methods were chosen from inventions during study 1 (with minor modifications, to improve comprehensibility and flow, see

Figure 13). The selected methods reflect common designs and conceptual errors (such as variability equals central tendency). All methods were well defined, though not all of them were applicable to all contrasting cases (for example, one method suggested that spread can be measured using Mode. However, not all contrasting cases included repeating values). The methods were given in a paper booklet. Students were asked to evaluate the methods using the Invention Lab. When a method was found successful, students were asked to evaluate the same method on a new set of contrasting cases. When a method failed, students were instructed to make a note of that and attempt the subsequent method on the same contrasting cases.

If evaluation (on top of intuitive ranking) is sufficient, then this condition should be at least as productive as the Full Invention condition. Furthermore, this condition may be better than Full Invention, given that these students are evaluating methods that target specific conceptual errors.



Methods Bank

In the next pages you will see methods that were created last year by students in your school. The purpose of the methods is to calculate the closeness of the points in each graph.

Each method should give a single number, a score, for closeness. Your goal is to evaluate these methods. Do they work? A good method should always choose the better trampoline.

Please write below each method whether it works or not.

- If a method works for the two trampolines, then submit it and answer the questions below it. Then try to apply it again to the next trampolines.

- If a method does not work, delete it from the interface ("delete all") and try the following one.

Good luck!

13. Half Minus Half

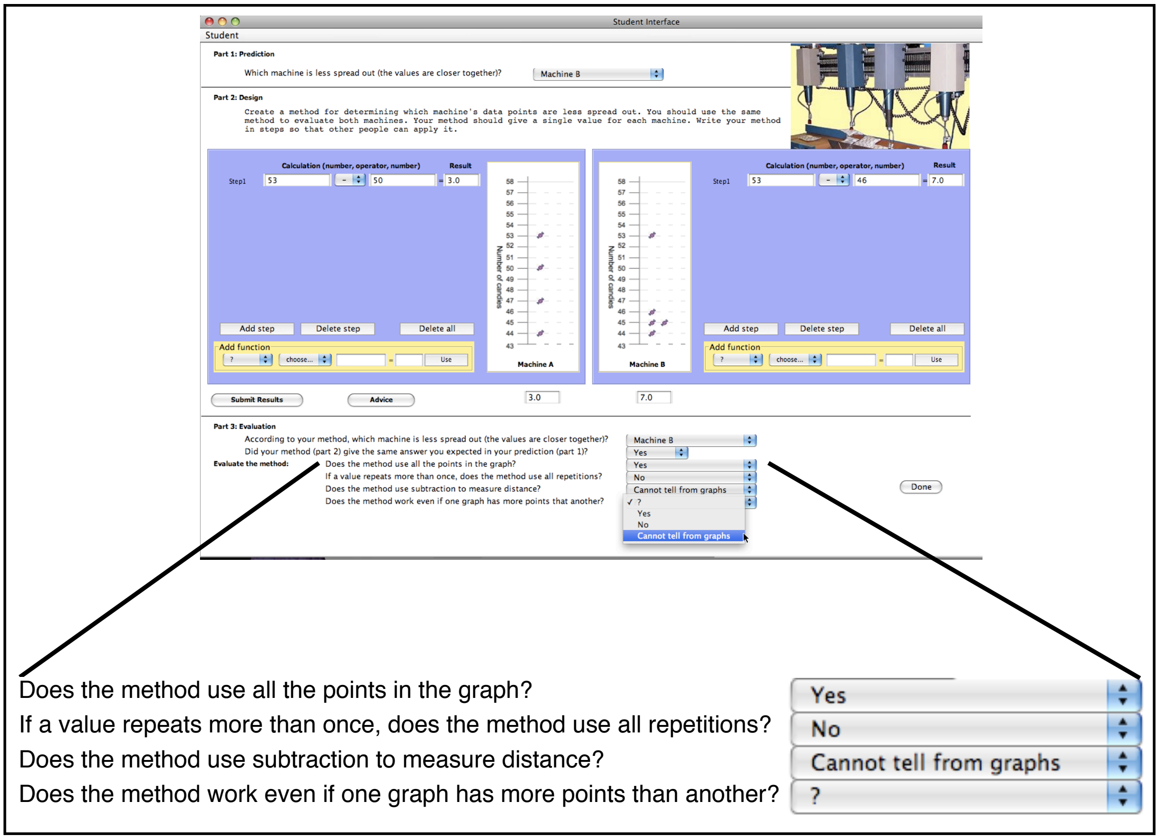
1. You add together all the points in the higher half.
2. You add together all the points in the lower half
3. You subtract the lower half from the highest half

Figure 13: An example for a pre-designed method given to students in the evaluation conditions in Study 2. On top - instructions to students. On the bottom - an example for a method. Students were asked to evaluate this and other methods using the Invention Lab. The methods were given in a paper booklet.

3. Reflection Support: The Reflection Support condition resembled Method Evaluation, with the addition of a reflection phase during evaluation of successful methods. In that phase, students were asked to answer reflection questions about the deep features of their method (Figure 14). These questions were extracted from the cognitive model of the Invention Lab: Does your method use all numbers? If a value repeats more than once, does it use all repetitions? Does the method use subtraction to measure distance? And does your method work for graphs with different number of points? This form of self-explanation was designed to help students internalize the deep features of the domain, as seen in the corrective self-explanation literature (c.f., Siegler, 2002). This condition targeted H2: The Domain Knowledge hypothesis. By evaluating the methods using an explicit set of principles, students may generalize more across methods and contrasting

cases. Also, the hypothesis suggests that noticing features is a key aspect of IPL. In that case, having students reflect on critical features of methods may promote better (and more efficient) learning than if students must stumble into these features while struggling to make their own inventions.

Within the invention cycle, reflection items appeared after students had a chance to evaluate their own methods. Also, students reached this step only after finding that a method had worked for a specific set of contrasting cases. Asking students to reflect only then had several goals. First, a reflection during the problem solving itself tends to impose too much cognitive load (Gama, 2004; Roll et al., submitted). The approach taken here encourages students to analyze what made the method work, and thus may achieve the effect of menu-based self-explanation (Aleven et al., 2006). Due to technical glitches and pressing timeline, the reflection questions did not undergo extensive user testing and thus may not represent that an ideal implementation of this “deep feature focusing” manipulation (c.f., Butcher & Aleven, 2008).



Part 1: Prediction

Which machine is less spread out (the values are closer together)? Machine B

Part 2: Design

Create a method for determining which machine's data points are less spread out. You should use the same method to evaluate both machines. Your method should give a single value for each machine. Write your method in steps so that other people can apply it.

Step1 53 - 50 = 3.0

Step1 53 - 46 = 7.0

3.0

7.0

Part 3: Evaluation

According to your method, which machine is less spread out (the values are closer together)? Machine B

Did your method (part 2) give the same answer you expected in your prediction (part 1)? Yes

Evaluate the method:

Does the method use all the points in the graph? Yes

If a value repeats more than once, does the method use all repetitions? No

Does the method use subtraction to measure distance? Cannot tell from graphs

Does the method work even if one graph has more points that another? ✓ ? Yes No Cannot tell from graphs

Does the method use all the points in the graph? Yes

If a value repeats more than once, does the method use all repetitions? No

Does the method use subtraction to measure distance? Cannot tell from graphs

Does the method work even if one graph has more points than another? ?

Figure 14: The Invention Lab Reflection Support. On top, the overall lab interface. On the bottom, the reflection questions. The possible answers to all questions were identical and are shown above. The features in the questions are the main target features from the cognitive model of the Invention Lab. Feedback on answers was not provided.

The three conditions differed also with regard to the motivational aspect of IPL. While students in the Full Invention condition had high agency over the solution process, students in both evaluation conditions did not design their own methods. Table 17 summarizes the three conditions.

Table 17: Experimental conditions in study 2

<table>
<tr><th></th><th>Full Invention</th><th>Method Evaluation</th><th>Reflection Support</th><th>Full Invention in study 1</th></tr>
<tr><td rowspan="3"><b>Intuitive ranking:</b></td><td colspan="4">Evaluation of contrasting cases with regard to the target concept</td></tr>
<tr><td colspan="3">Adaptive contrasting cases target previously expressed conceptual errors</td><td>Generic contrasting cases</td></tr>
<tr><td colspan="3">Immediate feedback on intuitive ranking</td><td>Feedback during class discussion</td></tr>
<tr><td rowspan="2"><b>Design:</b></td><td>Design of mathematical methods</td><td colspan="2">Students receive pre-designed methods</td><td>Design of mathematical methods</td></tr>
<tr><td colspan="3">Students apply the methods in steps using mathematical expressions</td><td>Methods are unconstrained</td></tr>
<tr><td rowspan="2"><b>Evaluation:</b></td><td colspan="3">Immediate feedback on evaluation</td><td>No feedback on evaluation</td></tr>
<tr><td colspan="2"></td><td>Analysis of method using given features</td><td></td></tr>
</table>

5.2 Methods

(a) Design

The study focused on the second topic from study 1, Variability. Students in all conditions used the Invention Lab for their invention tasks. Students also used a tutoring system (built for the study) during practice (see Figure 15). Tests were given on a paper, and I provided the in-class instruction. Similarly to Study 1, instruction focused on Mean Absolute Deviation. Unlike previous IPL and other similar studies (c.f., Kapur & Lee, 2009; Roll, Aleven & Koedinger, 2009; Schwartz & Martin, 2004), students within the same classes were randomly assigned to conditions.

Students were aware of the different conditions, as students in some conditions received a booklet while others received only a single instructions page. Therefore, I explicitly told students that not all of them are going to do the same activities, and briefly explained that some students would design while other evaluate the methods of their peers from the previous year.

(b) Participants

The study was conducted in the same school as study 1 with the same grade level (7th grade), taught by the same teachers. Likewise, the study included 2 levels of classes: 3 regular classes (taught by one teacher) and 3 advanced classes (taught by a different teacher, see Table 18). Another class from the same cohort taught by a different teacher was used as a pilot, and its data is not included in this analysis.

Table 18: Participants in study 2. Overall students (Invention Lab teams)

	Regular classes	Advanced classes
Method Evaluation	14 (7)	16 (8)
Reflection Support	14 (8)	15 (8)
Full Invention	17 (9)	16 (9)

Students worked on the Invention Lab in pairs. In general students worked with the same partners during both days, though some exceptions were made due to absentees. Students could choose their own partners unless the teacher did not allow that specific match.

(c) Procedure

The study spanned two days with two periods per day (in most classes these were consecutive). Activities throughout the four periods were identical in all conditions (see Table 19). The first day began with a short introduction, followed by the Invention Lab tutorial. The tutorial was motivated using the same Trampoline cover story that was used later in the first invention task. Following the tutorial students used the lab to compare the averages of the first two trampolines. This was done to help students get accustomed to the lab during a relatively simple and well-defined task. After calculating the average (and noticing that it fails) students moved on to the first invention activity (which was defined by their condition). The invention cycle was limited by time, and students completed the contrasting cases at their own pace. The first invention task was concluded with a short whole-group discussion in which students across all three conditions shared their solutions and difficulties. The second invention task had a similar structure, and was divided across the two days. Following the summary discussion of the second task, students received about 7 minutes of direct instruction on a correct method for computing spread (MAD). The instruction was taken almost verbatim from study 1. Students then worked on the practice tutor in which they solved problems where they needed to compute MAD and identify valid inferences from given data (see Figure 15). Students were given sufficient time, about 25 minutes, to solve all 10 problems in the practice environment. The study was concluded with a post-test.

Table 19: Procedure of study 2

Day	IPL		
Day 1	Pre-test and introduction		(19 min)
	Invention Lab tutorial		(15 min)
	Invention task 1: The Olympic Trampoline	Problem Setup	(10 min)
		Invention activity (varied by condition)	(23 min)
		Discussion	(3 min)
	Invention task 2: Candy Packages	Problem setup	(3 min)
		Invention activity	(28 min)
Day 2		(varied by condition)	
		Discussion	(3 min)
	Instruction		(7 min)
	Practice		(25 min)
	Post-test		(32 min)

(d) Materials

Invention tasks

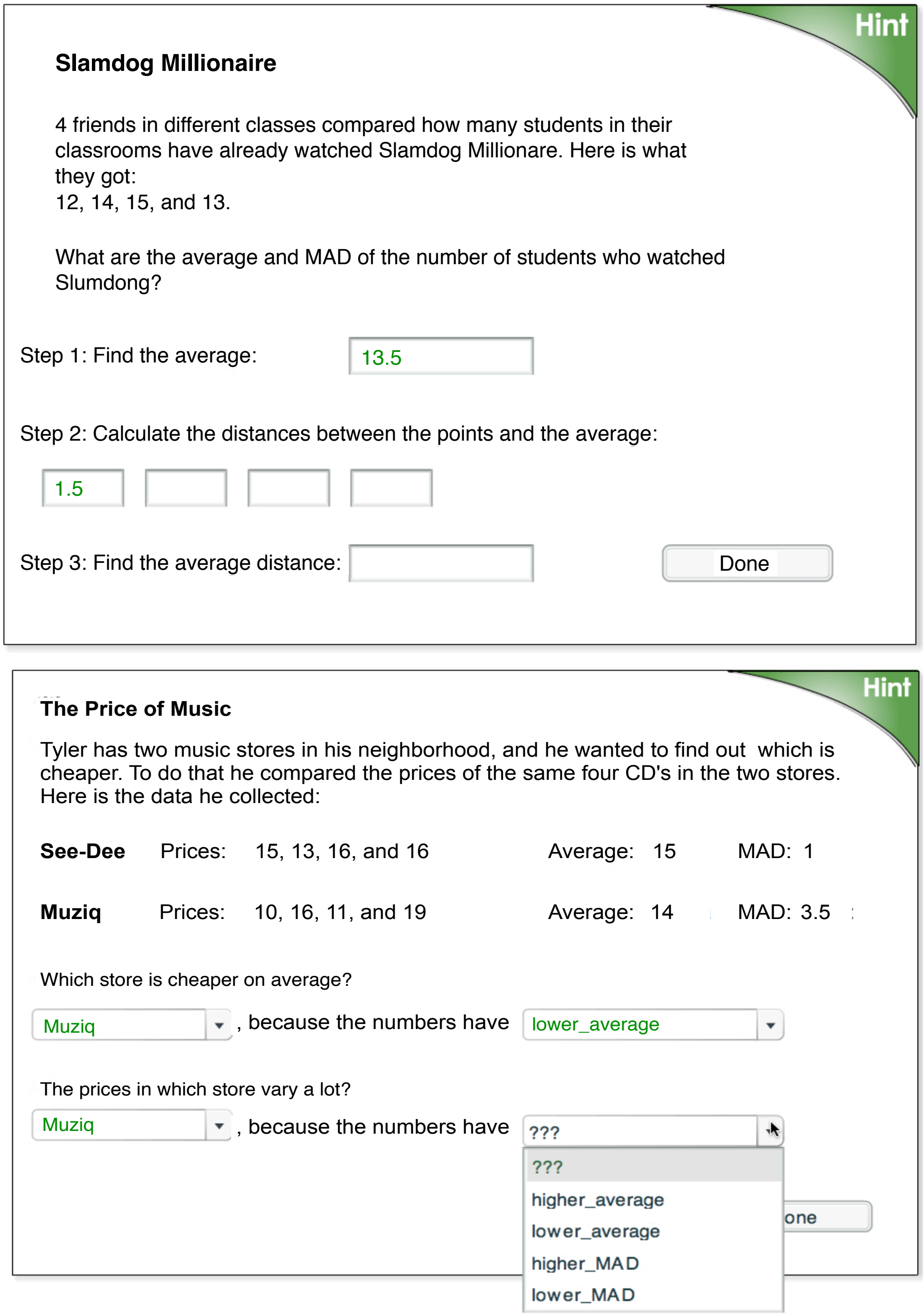
The study involved two invention tasks. The first asked student to find which trampoline is more fair to all athletes, and was essentially identical to the first invention task in Study 1. The second invention task used data about number of candies that machines pack into packages in a candy factory (called KanD). Students were asked to find the machine that makes more or less the same size of packages. The values in the Trampoline task were between 0 and 15. The values in the KanD problem were between 43 and 58.

The first set of contrasting cases in the Trampoline task was very basic and emphasized the range of the data (see Figure 1). The first set of contrasting cases in the KanD problem assumed that students are more proficient and thus differed on several parallel dimensions such as repeated values, sample size, and symmetry (44, 47, 50, 53 vs. 44, 45, 45, 46, 53). Subsequent contrasting cases in both cases were constructed on the fly based on students' inventions.

Show-and-practice

The show-and-practice component included a PowerPoint presentation given by me. The presentation resembled the instruction in study 1. It included a graphical, verbal, and mathematical explanation of the procedure for finding MAD, together with some guided practice and fading scaffold. It was followed by a practice.

Students practiced the taught material using a tutor for MAD, and this time worked individually. The tutor was built using CTAT (in Example Tracing mode; Aleven, McLaren, Sewall & Koedinger, 2008), and resembled a conventional coached practice tutoring environment, with detailed scaffolds, fix sequence of steps, well-defined answers, and immediate feedback. It included three units (see Figure 15). The first unit asked students to state the steps in finding MAD (by choosing the right steps from drop-down menus), and to apply these to new sets of trampolines (2 problems, 4 data sets overall). The second unit included similar problems using different cover stories, and without reiterating the formula for finding MAD (2 problems, 2 data sets overall). The third unit included conceptual problems that focused on the different uses of average and MAD, as well as grounding the meaning for high vs. low MAD (6 problems).



Slamdog Millionaire

4 friends in different classes compared how many students in their classrooms have already watched Slamdog Millionare. Here is what they got:
12, 14, 15, and 13.

What are the average and MAD of the number of students who watched Slumdong?

Step 1: Find the average: 13.5

Step 2: Calculate the distances between the points and the average:

1.5

Step 3: Find the average distance:

The Price of Music

Tyler has two music stores in his neighborhood, and he wanted to find out which is cheaper. To do that he compared the prices of the same four CD's in the two stores. Here is the data he collected:

See-Dee	Prices: 15, 13, 16, and 16	Average: 15	MAD: 1
Muziq	Prices: 10, 16, 11, and 19	Average: 14	MAD: 3.5

Which store is cheaper on average?

Muziq , because the numbers have lower_average

The prices in which store vary a lot?

Muziq , because the numbers have ???

Figure 15: The practice tutor included two types of practice: procedural (on top), in which students applied the MAD procedure learned in class, and conceptual (on bottom), in which students evaluated inferences from given MAD and Average of different data sets.

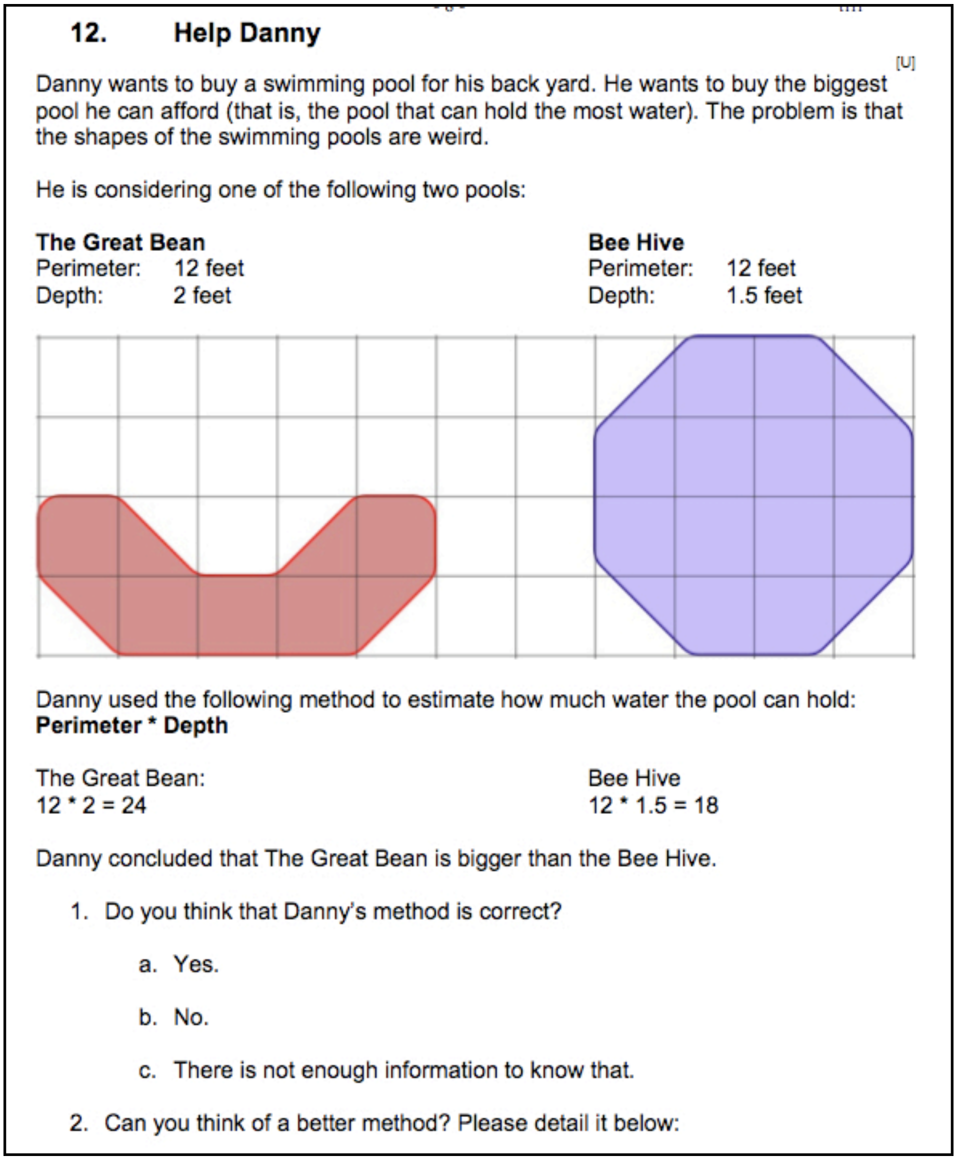
Assessment materials

The isomorphic test items were isomorphic to questions from the practice tutor, albeit given on a paper. These items assessed procedural fluency and conceptual understanding.

New-strategy items included items with and without embedded learning resource, counterbalanced between forms. The effect of IPL on new-strategy items in study 1 was somewhat different than the effect reported by Schwartz and Martin (2004). While Schwartz and Martin reported improved performance on new strategy items with learning resource, study 1 found improved performance on items without learning resource. I explained this discrepancy by suggesting that the new-strategy items in study 1 were closer to the show-and-practice instruction and thus could be solved more easily without learning resources (on the graphing post-test). To evaluate this explanation, new-strategy items in this study were designed to be almost impossible in the absent of a learning resource. One new-strategy item required students to divide MAD by average in order to find the relative magnitude of the MAD. The other new-strategy item required students to ignore small errors when calculating MAD.

A forth type of item, debugging, asked students to identify errors in faulty methods.

The study also targeted the general invention ability of students. Invention ability was assessed by giving students a faulty formula for volume, one that generates a wrong prediction when applied to the given contrasting cases (see Figure 16). Students were asked to evaluate the given formula (which was perimeter times depth, rather than area times depth), and were asked to think of a better one. The topic of the problem was a new one, volume, to evaluate whether students could isolate acquisition of strategic knowledge of invention strategies from acquisition of domain knowledge. Volume was also chosen under the assumption that it would have a sufficient relevant domain-knowledge base to build upon (which was probably incorrect in retrospect).



12. Help Danny

[U]

Danny wants to buy a swimming pool for his back yard. He wants to buy the biggest pool he can afford (that is, the pool that can hold the most water). The problem is that the shapes of the swimming pools are weird.

He is considering one of the following two pools:

The Great Bean		**Bee Hive**	
Perimeter:	12 feet	Perimeter:	12 feet
Depth:	2 feet	Depth:	1.5 feet

Danny used the following method to estimate how much water the pool can hold:
Perimeter * Depth

The Great Bean:
12 * 2 = 24

Bee Hive
12 * 1.5 = 18

Danny concluded that The Great Bean is bigger than the Bee Hive.

1. Do you think that Danny's method is correct?
   a. Yes.
   b. No.
   c. There is not enough information to know that.
2. Can you think of a better method? Please detail it below:

Figure 16: Invention test item. This item, in the domain of volume, has a similar structure to the variability invention tasks, albeit in a different domain. Students have learned to calculate volume prior to the study in their regular classes.

Last, the study included a self-report survey. The survey included the self-efficacy items from study 1 (adapted from Pintrich 1990, 1993), in addition to personal and situational interest items adapted form Mitchell (1992).

5.3 Results

This section first describes students' behavior while using the Invention Lab, followed by an analysis of students' performance on the tests and the practice tutor.

(a) Working with the Invention Lab.

Log files from the study were analyzed to improve our understanding of the invention process, to identify differences between conditions, as a manipulation check, and to evaluate whether the Invention Lab supports the invention activity sufficiently. Students worked with the lab in pairs (see Table 20). Each team performed on average 441 actions using the Invention Lab. While Reflection Support students performed fewer actions overall, this effects is only marginally significant and will be explained later (Full Invention: 477; Method Evaluation: 461; Reflection Support: 379; *F*(2,39)=2.8, p<.08). Given that Full Invention students need to come up with methods on their own, it is surprising that they do not have fewer actions. The logs show that 80% of students' actions matched the defined task progression (intuitive ranking -> design -> evaluation) or the input process (value -> operator -> value). Several factors appear to account for the 20% that were out of sequence. First, students may have performed the wrong actions intentionally, either because they did not understand basic attributes of the task (e.g., a method that does not generate the correct prediction should be revised), or because they tried to game the system (e.g., by clicking 'done' when a problem was too challenging, Baker et al., 2008). The system may also be to blame for some of these out-of-sequence actions, either because of unintuitive interface features (for example, unintuitive order of operations to use a formula), or because of various implementation issues (mainly slow response time due to the large cognitive model). In addition, this rate may simply be an outcome of the novel interface. There was a significant effect for condition on the rate of out-of-sequence actions, controlling for class level (*F*(2,42)=3.8, p<.04). Analysis of the contrasts shows that Reflection Support had a higher rate of out-of-context actions (24%) compared with either Full Invention (19%) or Method Evaluation (19%). The higher rate of out-of-sequence actions in the Reflection Support condition is probably due to the additional interface elements (the reflection questions, see details below). The system, in response to out-of-context actions, refocused students on the next action or task element to be completed. Since this analysis focuses on invention behavior rather than the interface itself, out-of-context actions are excluded from subsequent analysis.

Table 20: Number of teams working with the Invention Lab

Condition	# of teams regular, advanced	# of teams that voluntarily switched conditions	# of actions per team (% of out-of-context actions)
Full Invention	9, 9	-	477 (19%)
Method Evaluation	8, 8	4	461 (19%)
Reflection Support	7, 8	2	379 (24%)

How do inventions look like?

Analysis of students' inventions with the Invention Lab shows that the lab supports progression in students' conceptual understanding of the domain. Table 21 shows a typical sequence of methods invented by one team on one set of contrasting cases (Full Invention condition, regular class, trampoline problem). We join the team after they already solved the first set of cases using range. The lab presented them a new set of cases, (1,3,5,7,9) vs. (1,4,5,6,9), and students rank the cases correctly (Trampoline B has less spread). The students begin by applying the previously successful method, range, and find that it fails on the current set of cases (actions 1-13). They move on to attempting multiple central tendency methods (actions 14-33), and then try range again (34-42). Then they decide to extrapolate range to the second-furthest pair of numbers in each method. They first list the two distances without connecting them (43-61), then subtract the two distances (applying the concept of distance to the distances themselves), but notice that it gives the opposite result (actions 62-90). To address this, they decide to reverse all operators - and use addition instead of subtraction (actions 91-176). When this fails too, they go back to distance, only that this time they add the distances rather than subtracting them (177-200). This method is found to be successful, and the lab presents the students with two new contrasting cases. In addition to students' progression at the domain level, the students came across rich experiences also at the meta-level of scientific inquiry behavior. At first they did not self-evaluate their methods before submitting them to the system, and gradually they began evaluating these prior to system feedback. They encountered the limitations of methods that do not give a single number. The students also experienced that the same method always gives the same result when applied to the same data.

Table 21: Typical interaction with the Invention Lab

Time m:ss (Action #)	Method for Trampoline A (data: 1,3,5,7,9)	Method for Trampoline B (data: 1,4,5,6,9)	Outcome	System feedback
0:53 (10)	9 - 1 = 8	9 - 1 = 8	Students submit method, then reply that its outputs match their intuitive ranking	"But your answer in the last question is not the same as your initial prediction."
1:09 (13)			Students revise evaluation, and say method does not work	"Then please go back to Part 2: Design and revise your method."
3:09 (27)	Median(all) = 5	Median(all) = 5	Students notice that method does not work prior to submission and delete it	
3:28 (33)	Mean(all) = 5	Mean(all) = 5	Students notice that method does not work prior to submission and delete it	
4:16 (42)	9 - 1 = 8	9 - 1 = 8	Students submit method but delete it before beginning evaluation	
7:35 (61)	9 - 1 = 8 7 - 3 = 4	9 - 1 = 8 6 - 4 = 2	Students submit method, though it does not assign a single value to each graph	"Please involve each of your steps in the final step you use. In the left graph, step 1 is not used to compute the final result."
10:54 (90)	9 - 1 = 8 7 - 3 = 4 step1-step2=4	9 - 1 = 8 6 - 4 = 2 step1-step2=6	Students submit, say that method does not work (since B should have less variability)	"Then please go back to Part 2: Design and revise your method."
17:14 (176)	9 + 1 = 10 7 + 3 = 10 step1+step2=20	9 + 1 = 10 7 + 3 = 10 step1+step2=20	Students submit method but delete it before beginning evaluation	
19:03 (200)	9 - 1 = 8 7 - 3 = 4 step1+step2=12	9 - 1 = 8 6 - 4 = 2 step1+step2=10	Students submit method, confirm that it works, and move on to the next set of cases.	"Good. Click Done to continue"

This sequence also demonstrates some limitations of the lab. The overall process as described above required about 200 actions. Many times throughout the process the students got confused and did not understand what is expected of them. Perhaps the worst example is that students entered 'x' instead of '-' during the third minute. They attempted to fix it right away, but due to limitations of CTAT widgets, this command did not get to the cognitive model. The discrepancy between the methods as perceived by the students vs. the system was not fixed until almost 7 (!) minutes later. Luckily, since the system does not often give feedback, this discrepancy did not interfere much with the invention process. (One should keep in mind that off-task time is also an occasional feature of normal classrooms, where students may mind-wander during lecture, and of reform classrooms, where project-based groups can get stuck or discuss the weekend.)

A quantitative analysis of the logged data reveals interesting characteristics of inventions, as summarized in Table 22. Due to a bug in the logging code of the Invention Lab, its interpretations of students' methods were not logged. However, the wealth of data offers other interesting analyses. One relevant parameter is the complexity level of students' methods. Since inventions in the Invention Lab are made of steps, the number of steps in a method can serve as an indicator for the method's complexity. Note that the number of steps does not include further complexity that is introduced by using functions (such as average, count, or sum). The average depth of a method was 1.7 steps (depth of a method is defined as the number of steps a method has when either submitted or deleted). While this may seem low, this number is actually within the roam of expectations. First, even single-step methods can be fairly complex (see Figure 13). Second, this matches many of the methods previously seen in Study 1. Last, this includes students who began a certain method and scrapped it off the board before completion. While not all methods need to be complex, the lab should support the authoring of more complex ones. The maximum number of steps in methods, averaged across teams, was 3.7, and the median was 3 - that is, at least half of the teams created at least one method with 3 steps or more. There was no significant difference between conditions.

Not all methods were submitted. On average, students revised 66% of their methods (that is, deleted the whole or part of the method) before receiving any feedback from the system. There was no effect for condition. Such high voluntary revision rate may suggest that students are good at evaluating their own methods. However, this is not necessarily the case, since students had to revise 60% of their submitted methods as well. That is, only 40% of their submitted methods completed the specific invention cycle and led to an introduction of new contrasting cases. Moreover, 22% of the submitted methods were inconsistent, that is, different methods were applied to both contrasting cases. It may be that the students were submitting methods they knew would fail, in the hope that the system will not notice that, or in the absence of better alternatives. Alternatively, students preferred using the scaffolded evaluation rather than do so mentally prior to submission.

Effect of class level

Interestingly, the invention process was similar across class levels. For example, students in both levels had a similar number of steps on average and in submitted methods. Students also spontaneously revised a similar portion of the methods prior to receiving feedback. However, it seems that advanced students made better domain-specific decisions when performing these actions. In other words, while the invention process was similar, its outcomes were different. For example, only 50% of the methods submitted by teams in the advanced classes had to be revised, compared with 73% in the regular classes (*F*(1,38)=16, p<.0005). Also, advanced students worked faster, with an average of 11 seconds per action, compared with 15 seconds per action in the regular classes (*F*(1,43)=4.0, *p*=.05). These differences between levels allowed the advanced teams to complete twice as many contrasting cases within the same time frame (advanced teams completed an average of 9.6 contrasting cases, regular teams did only 4.2; *F*(1,35)=9.6, p<.01). It may be that advanced students attempted more methods using mental math and implemented only the more successful ones.

This number of contrasting cases completed by students in advanced classes is relatively high. In study 1, students completed about 4 sets of contrasting cases in approximately the same time (this number is only an estimate), similar to the number of contrasting cases completed by the regular classes in this study, and much lower than the advanced teams. The introduction of adaptive contrasting cases is not likely to contribute to the increase in these numbers, since the tailored contrasting cases target students' difficulties, and thus are not likely to be solved easily. However, the tailored cases introduce features gradually, and aid students by basing current contrasting cases on previous ones. Students in the advanced classes may have been able to take advantage of this. Perhaps a more likely explanation for the apparent time difference in case exploration and invention across studies, is that students in the Invention Lab were facilitated in method design and method implementation by interface scaffolding that prompts for designing a mathematical method (as opposed to intuitive or graphical methods students attempted in Study 1) and that provides hints for the kinds of mathematical operations that might be performed. Another possible time-saver is the fact that the Lab computes the mathematical operations for students and saves their time and energy to focus on features of the method and its success.

Table 22: Main findings in the Invention Lab data

Aspect of invention	Claim	Support in data
Invention behavior	Students have some intuitive understanding of the scientific reasoning process, however, they are in need for more explicit support and feedback.	• Students revised spontaneously 2/3rd of their methods. Still, 60% of the submitted methods did not lead to successful completion of the cases. • 22% of the submitted methods were inconsistent, that is, different methods were applied to both cases.
Effect of class level	Students in both class levels had similar invention patterns. Students in the advanced classes made better decisions that resulted in better methods.	• There were no significant differences between conditions with regard to the average length of submitted method and ratio of deleting methods to all actions. • Students in advanced classes had a higher rate of successful methods (among submitted methods), they acted faster, and completed more than twice as many contrasting cases.
Influence of design on process	Full Invention teams have a better match between their methods and the contrasting cases. Overall methods are similar.	• Full Invention students were able to successfully complete significantly more contrasting cases. • A trend in the data suggests that they revised fewer methods prior to submitting. • Their successful methods were shorter than successful methods in the other two conditions.
Effect of reflection questions	The reflection support component, as implemented in the study, led to confusion	• Students avoided these items (even though they answered the previous items). • Students overall completed fewer contrasting cases than the other two conditions. • Reflection Support students had a higher rate of out-of-context actions.

The influence of design on the invention process.

At the same time, we would expect to find differences in the design process. Specifically, while Full Invention teams targeted their methods at the contrasting cases, teams from both evaluation conditions applied random methods that may not order the two given cases correctly, that is, as identified during the intuitive ranking phase. A trend in the data suggests that Full Invention students deleted their methods less often. Full Invention students revised 63% of their methods prior to submission, while Method Evaluation and Reflection Support revised 68% of their methods prior to submission). This trend does not reach significance (*F*(2,41)=4.3, p<.2).

While the average length of methods was similar in all conditions, successful methods in the Full Invention condition were much shorter than their counterparts. Full Invention had an average of 1.1 steps in a successful method, Method Evaluation had 2.3, and Reflection Support had 2.0. (*F*(2,18)=12.4, p<.0005). Analysis of the contrasts shows that Full Invention differs from each of the two other conditions. There is no statistical difference between Evaluation Only and Reflection Support. Notice that successful methods in both evaluation conditions were longer than their average attempted method, while successful methods in the Full Invention condition were much shorter than the average. In fact, 93% of the successful methods submitted by Full Invention teams had exactly one step. Since the average length of methods was identical across conditions, it seems that Full Invention teams did not attempt to create short methods - rather, they only succeeded when using short methods. Alternatively, given the demand in the other conditions to use given methods that might not fit the cases, they are more likely to apply longer methods to cases that don't need them. Invention students, on the other hand, are unlikely to unnecessarily create longer methods without a good reason (i.e., when the cases don't require it).

Another difference between conditions is the number of contrasting cases students were able to complete. Full Invention students completed 10.4 contrasting cases on average, compared with 7.5 and 3.4 for Evaluation Only and Reflection Support respectively (*F*(2,34)=5.5, p<.01). An analysis of the contrasts shows that Full Invention condition differed from each of the other two, which were not statistically different from each other. This difference is essentially a manipulation check because only the Full Invention students were expected to complete cases whereas the others were supposed to evaluate whether or not methods worked using the cases. While not surprising this difference may be important with respect to potential differences in learning outcomes between the two groups.

The effect of the reflection questions

Reflection Support students had an additional set of questions to answer after submitting and evaluating their methods and prior to receiving new cases. A quick look at the data shows that students indeed answered the reflection questions. It also seems that the students took these questions seriously. They spent on average 8 seconds per question, and there was no single pattern of responses (that is, choosing 'yes' or 'no' on all questions).

However, other signs suggest that this support did not achieve its goal. Instead, there is evidence that suggests that the reflection questions caused confusion. Earlier I reported the rate of submissions to successful submissions, that is, submissions that led to new sets of contrasting cases. This ratio is significantly lower for Reflection Support students: 29% vs. 44% and 43% for Full Invention and Method Evaluation respectively ($F(2,37)=2.8$, $p=.07$). Analysis of the contrasts shows that the Reflection Support teams differed significantly from Method Evaluation and marginally significantly from Full Invention. The latter two are statistically indistinguishable. An interesting question is what happened with the 71% of methods that were submitted by Reflection Support teams but did not lead to successful conclusion. The reflection questions appeared right after students were asked to evaluate their own method. However, on 56% of the times in which students evaluated their methods, they chose to go back and revise these rather than move on to the reflection questions. One possibility is that students tried to avoid these questions rather than answer them. This apparent avoidance is especially striking given that no feedback was given on reflection questions. Another sign of confusion is students' percentage of out-of-context actions (that is, actions that do not match the flow of the task either at the inquiry level, e.g., evaluating before applying, or at the mathematical level, e.g., not using operators between values). As reported above, the rate of out-of-context actions was significantly higher for Reflection Support students than in the other two conditions.

Given that students spent time on the reflection questions and did not appear to answer them randomly, another possibility is that upon reading these questions students felt they ought to go back and revise their methods, perhaps to better match these features. However, there is no direct evidence that the reflection questions led to mental of reflection. Other measures of Reflection Support teams actions outside the reflection phase, including average time on action, were statistically identical to that of Method Evaluation. Thus, there is no clear evidence of prompts for reflection spilling over into mental reflection during the other phases. However, students did spend some time on these questions: On average 32 seconds for the 4 questions, which is about 2-3 times the reading time needed (using 250 ms per word) and arguably students don't need to keep rereading these questions to answer them for each case comparison. Analysis of correctness of student answers remains for future work.

(b) Practice environment

Students' performance in the practice tutor was logged and analyzed. Due to a technical error, data from one of the regular classes was not logged during practice. This reduced the statistical power (data was logged from only 8 students in each of the no-design conditions in the regular classes). To account for that, information from the two no-design conditions was merged for statistical purposes. Figure 17 shows students' success rate in the practice tutor. ANOVA of the success rate (correct actions out of all actions) as a function of condition in both class levels found a marginally-significant advantage for Full invention in the regular classes, (*Full Invention: 66%, Method Evaluation: 56%, Reflection Support: 62%; F*(1,25)=2.9, *p*=.1). There was no difference in the advanced classes.

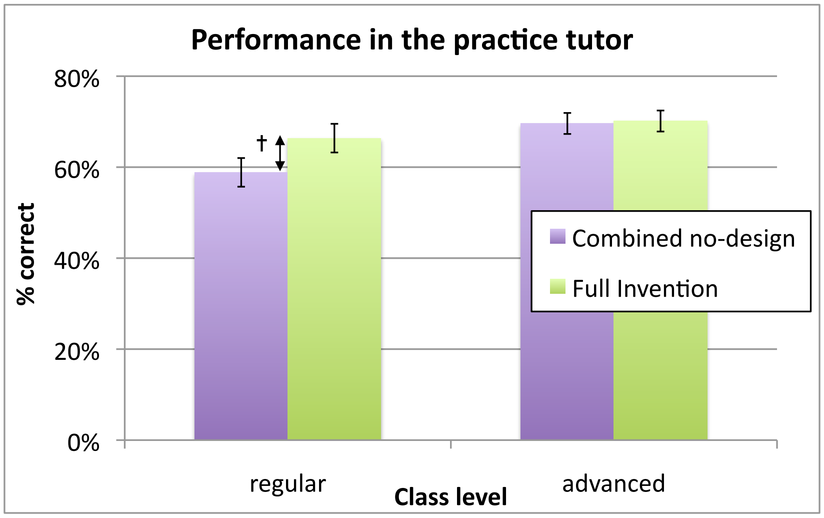


Figure 17: Success rate in the procedural practice environment. Due to lost data and small statistical power, data from both no-design conditions was combined for statistical purposes.

(c) Learning outcomes

Before the study began students were given a short pre-test, which included questions about variability and spread. Students in all three conditions (and two classroom levels) performed statistically similar to each other, and not significantly better than chance. Overall, students' performance improved from pre- to post-test (variability items: from 23% to 58% across conditions and class levels; *t*(90)=7.0, p<.0005; conceptual spread items: from 28% to 38% across conditions and class levels; *t*(90)=2.5, p<.02). Being at chance, the pre-test is not a significant predictor of performance during the study or the post-test, and thus is not used in further analysis.

Isomorphic measures

Table 23 includes a summary of the results of study 2. Isomorphic measures include procedural and conceptual items that are isomorphic to items practiced in class. There was no significant effect for condition on these items in either class level.

Table 23: Summary of results in study 2: score (SD).

Assessment	Regular classes			Advanced classes		
	Method Evaluation	Reflection Support	Full Invention	Method Evaluation	Reflection Support	Full Invention
Isomorphic items	.39 (.20)	.45 (.26)	.43 (.19)	.61 (.26)	.55 (.26)	.66 (.22)
Debugging items	.25 (.33)	.32 (.32)	.38 (.33)	.47 (.39)	.30 (.32)	.50 (.26)
New strategy with learning resource	.36 (.50)	.43 (.51)	.35 (.49)	.56 (.51)	.57 (.51)	.81 (.40)
New strategy without learning resource	.07 (.27)	.14 (.36)	.12 (.33)	.19 (.40)	.07 (.27)	.25 (.45)
Transfer invention task (topic: volume)	.29 (.47)	.36 (.50)	.18 (.39)	.25 (.45)	.20 (.41)	.13 (.34)

Comparing performance of IPL students in studies 1 and 2 is of interest, since both dealt with the same topic, time frame, background (school, teachers) and population. The two procedural items in study 2 were isomorphic to items given in study 1. Assuming that the problems were of equal difficulty, and that the samples draw from the same population, there are no significant differences between performance of students in study 2 to that of students in either condition in study 1.This evidence for “do no harm” is a reasonable goal for the first field-based evaluation of a new technology of this complexity.

New-strategy items

Since analysis of study 2 focuses on the role of design, and since no major differences were found in the process or outcome measures of both no-design conditions (Method Evaluation and Reflection Support), analysis of the test results in study 2 combines the data from students in Method Evaluation and Reflection Support conditions.

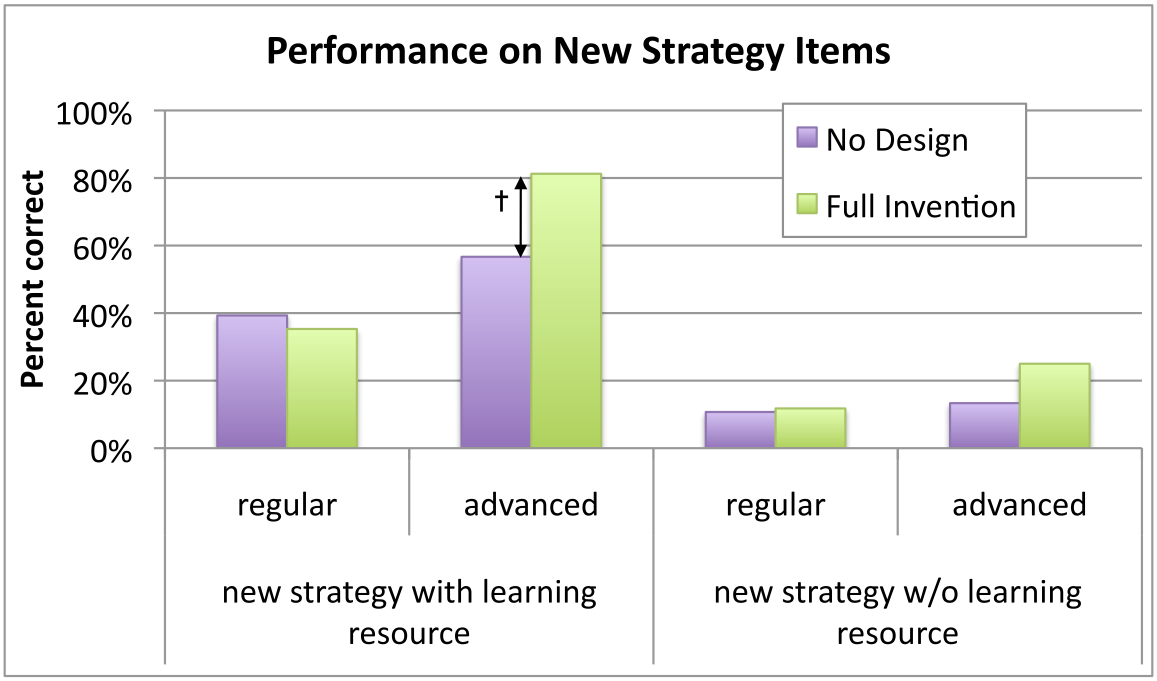


Figure 18: Performance on new-strategy items. Full Invention students in the advanced classes performed marginally-significantly better than their No Design counterparts on new-strategy items with learning resource.

The post-test included two types of new-strategy items: with and without embedded learning resource. Figure 18 shows performance on these items. Full Invention students in the advanced classes performed marginally significantly better than students in the No Design condition (81% vs. 57% respectively, *F*(1,44)=2.8, p=.1).

Unlike Study 1, and by design, performance on new strategy items without learning resource is almost at floor. There are no differences between conditions with regard to the types of errors students made.

Debugging

Three test items evaluated students debugging skills both within and outside the domain of statistics. Students needed to answer whether the procedure was correct, and if not, either improve it or locate the error. Figure 16 shows students' performance on the debugging tasks within the domain of variability (that is, students were asked to identify errors in two faulty variability formulas). Students in the Full Invention condition performed better than students in the combined No Design condition across levels (main effect for condition, *F*(1,88)=166.0, p<.05). In addition, advanced students in the Full Invention condition were the only group to perform above chance (*t*(15)=2.6, *p*=.02).

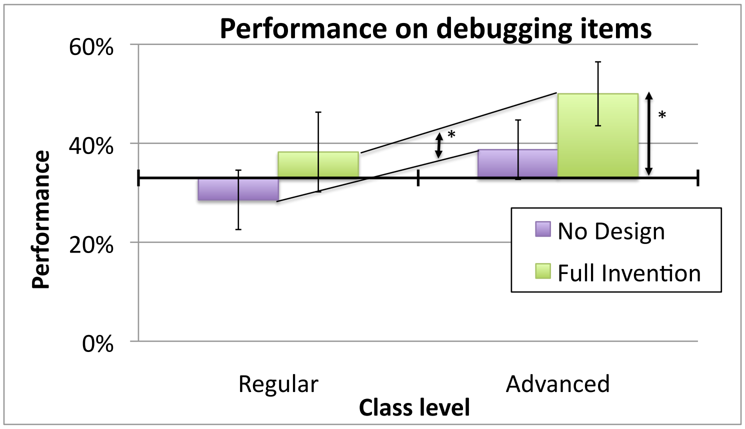


Figure 19: Students' performance on debugging items. Full Invention students in both class levels performed better than their No Design counterparts.

Transfer Invention Task

A different test item asked students to identify whether a method for calculating volume is correct, and if not, to debug it. This single item evaluated students' invention behavior in an unrelated domain. None of the conditions did better than chance at even recognizing that the given formula was faulty, indicating either that it is a poor item or that students did not have enough relevant prior knowledge (or experience) on this topic to engage in invention or a bit of both. There was no effect for condition or class level on students' performance on this item.

Motivation

The post-test included self-report surveys that measured self-efficacy, personal interest, and situational interest. No significant differences between conditions were found on any of these (see Table 24).

Table 24: Self report measures in study 2. No significant differences were found between conditions or from pre to post.

		Personal interest	Situational interest	Self efficacy	Effort during study
Method Evaluation	Pre-test	4.0	-	6.0	-
	Post-test	3.8	4.4	5.9	4.0
Reflection Support	Pre-test	3.7	-	5.9	-
	Post-test	3.9	4.3	5.8	4.0
Full Invention	Pre-test	4.4	-	6.0	-
	Post-test	4.4	4.6	6.1	4.0

During the study, I observed several teams who were engaged in activities that were not intended for them. More specifically, 4 teams in the Method Evaluation and 2 in the Reflection Support conditions were observed inventing methods (about 20% of the population in these conditions). These teams put down their booklet of pre-designed methods and discussed how to capture different properties of the data. Whenever I noticed that, I asked the student to resume evaluating pre-designed methods. At the same time, none of the Full Invention students was observed asking their peers to see or use their methods. No statistically significant differences (either on the process or outcome measures) were found between students who switched conditions voluntarily and the other students in their original condition. However, this voluntary invention suggests a possible motional benefit for the full IPL process.

Given that the assignment to conditions was done within classrooms, keeping track of the conditions of students who remained working during breaks was found technically challenging. Thus, this measure was not used in this study.

5.4 Discussion

(a) Summary of results

Study 1 provides evidence that the design and evaluation stages of Full Invention are important to better aid students in acquiring flexible knowledge relevant to new-strategy and debugging items (the intuitive ranking of contrasting cases alone appears to not be enough). Study 2 shows that design is a critical element. The study found that students who designed methods were better than students who evaluated pre-designed methods (with or without reflection support), on new-strategy and debugging items (albeit the effect is only marginally significant on new-strategy items). First, the effect on debugging items is especially striking, since these items were isomorphic to part of the invention task of the No Design students. These students were asked, during invention and during the test, to evaluate whether given methods are valid or not. Yet, they performed worse on these test items compared with students who designed their own methods. Another interesting observation is the identical pattern of results between studies 1 and 2. While study 2 seemed to have smaller statistical power, the benefits of the full invention activity were very similar to those found in study 1.

This study also found that Full Invention students in the regular classes performed marginally better in the practice environment. This effect appeared right after the classroom instruction, but did not carry over to the post-test. In that sense, this measure is a poor-man's version of future learning assessment - it evaluated students' ability to apply recently taught materials to new challenges. It may be that the additional practice during the practice stage eliminated these differences before the post-test. Analyzing the learning curves of students in the practice tutor would be of interest, but was not done yet.

Both study 1 and study 2 found behavioral evidence for the motivational benefits of IPL. In study 2 this materialized in students adopting activities designed for the Full Invention students but not vice versa. Unlike study 1, study 2 did not find motivational benefits using self-reports. This may be due to the novelty of the Invention Lab and the task. In study 1, the control-group students were engaged in practices that resembled their conventional class – that is, a frontal lecture in a class setting. In contrast, in study 2, students in the control condition were engaged in invention tasks, worked in groups, and used the Invention Lab. It may be that the novelty of these factors had a greater motivational effect compared with the differences between conditions.

The relative lack of statistical significance in study 2 can have several explanations. First, it may simply be because design is not a critical factor of IPL. The faulty predesigned methods may have been just as good. However, this explanation is not likely. The almost identical pattern in the results of study 2 and study 1 suggests that there is more to it. Lack of sufficient statistical power is always a potential explanation for null effects. Lack of power can occur due to small sample size (as indeed was the case with the practice tutor, due to lost data, too much variance in the data and lack of sufficient covariates to account for other sources of variance). It can also happen if the observed effect was smaller than anticipated (and smaller than in Study 1). There are several reasons why, if the design phase is indeed important for flexible learning. The observed effect may have been diluted due to the possible cross-condition contamination. Tutor messages may in fact have added to this. No design students were also getting the messages about how the method entered did not appropriately rank the cases and that they should try again, are thus may have been more likely to enter a method different from what was on paper. Another potential reason is that applying the pre-designed methods required transfer from paper to computer (and usually from one team member who held the paper to another who held the mouse). This may lead to design-like processes.

An important question is whether the Invention Lab is to blame for the results. In other words, was the Invention Lab successful in facilitating IPL, or did it fall short of achieving that? It is hard to judge the outcomes of the lab, given that this study did not include direct instruction condition. Students using the lab during the study and the pilots demonstrated their ability to invent using the lab. Moreover, these methods resembled the methods invented on paper and pencil during study 1, and there is evidence that students' thinking evolves while using the lab. Comparable pre-to-post gains between studies 1 and 2 also suggest that the lab achieved its foal. Still, several attributes of this process make it suspicious. First and foremost, the high ratio of out-of-sequence actions and other errors (made by the students or the lab) may have hindered the quality of students' reasoning. Of course, without the same level of process instrumentation in Study 1 or other prior IPL studies, it is hard to know whether such deviations from quality reasoning are just as (or even more) frequent in face-to-face settings. The high number of completed contrasting cases, plus the short length of methods and short time between actions, suggests that some students may have been more oriented toward performance goals than mastering goals. The lab may have changed the character of the task from a search for a global method to a hunt of local solutions.

Using the lab for the first time for such a short duration certainly imposes extrinsic cognitive load and steep learning curves. Using the lab over longer periods of time may resolve these issues, and with them reduce confusion, frustration, and out-of-context actions. The lab could certainly use some improvements, especially in encouraging students to connect across contrasting cases (and not only within the same set of cases). However, even in its first field experience, the Invention Lab helped students explore the tapestry of data analysis and scientific reasoning. The relative benefits of Full Invention students in study 2 were similar to the benefits of the corresponding students in study 1, which used paper and pencil inventions. These results, as well as data analysis with the lab itself, suggest that the Invention Lab was successful at achieving its goal.

(b) Reflection on the Reflection Support

Performance on the different post-invention measures did not find a difference between Reflection Support and its cousin, Method Evaluation. However, invention patterns of Reflection

Support suggest that the reflection questions may have had a negative effect on invention behavior. Certainly, the implementation of the reflection phase could be improved through further pilot testing.

The reflection questions were given after the students had invented methods that worked for the specific contrasting cases. They did not help students debug their methods simply because these methods were not in need for debugging. Furthermore, since the methods worked for these cases, it was hard for students to imagine their limitations. Poor design decisions were made not only with regard to the timing of the questions, but also their content. The contrasting cases gave students opportunities to experience the different features. The reflection questions, on the other hand, asked students about features that were not apparent in the data. Students at this level are not used to analyzing features of methods, and were in need for more support and better guidance. Since there was no feedback on students' answers to these questions, that potential opportunity was lost and students may have begun to not take the questions seriously.

(c) On scientific reasoning skills

Perhaps one of the clearer evidence that Study 2 supplies is that students lack basic skills and knowledge regarding the scientific reasoning process. During invention, 20% of the submitted methods were inconsistent (that is, different methods were applied to the different contrasting cases). 40% of the submitted methods were consistent but failed to rank the contrasting cases appropriately. Given that the average length of a submitted method was 1.7 steps, these high numbers are somewhat surprising. Students often resumed earlier faulty methods, especially ones that were based on central tendency. Students demonstrate poor understanding of the scientific method also when applying invention skills to isomorphic tasks in the same domain outside the lab. 45% of the students wrote that the buggy versions of MAD given to them during the test were valid, even though the task was isomorphic to the one in the lab - it featured contrasting cases with clear ranking, and the methods were already applied to the two cases.

Notably, there was no effect for class level or condition on these items. This is especially surprising for both no-design conditions, since their activity during the study was identical to the one in the test - to identify whether given methods are valid or not.

Chapter 6
General discussion: revisiting the research questions

6.1 Unpacking IPL

The first research question addressed by this thesis was to analyze, identify, and evaluate the critical elements of IPL. This thesis focused on the invention process, rather than other elements (such as collaboration, role of teacher, etc). The invention process can be mapped onto the scientific method, as defined by Popper (2002). This should come at no surprise. Invention, like other scientific endeavors, is an iterative process in which students construct methods to explain phenomenon they observe. More specifically, three phases were identified in IPL: Intuitive ranking, in which students intuitively evaluate the given data (or notice the phenomena they need to explain); Design, in which students design a mathematical procedure that, when applied to the data, should match their intuitive ranking (in more scientific terms, they construct a mathematical method that explains the phenomena); and last, evaluation, in which students evaluate their method against their intuitive ranking. Needless to say, scientific methods have more to them than invention. However, the invention activity emphasizes important scientific elements such as using data to form conclusions.

The thesis goes on to define other elements of IPL. Many of these were observed qualitatively during the different studies and still require hard-data to back them up. One of the more obvious elements is the contrasting cases, or the use of two sets of data that differ on one or more deep features of the domain. At the domain level, these cases are intended to direct students' attention to the deep features. At the metacognitive level, contrasting cases provide opportunity for self-monitoring in that intuitive ranking of the cases can serve as a baseline against which students can evaluate their inventions. At the motivational level, contrasting cases give room for incremental progress, and thus may improve students' self-efficacy. While at first contrasting cases should vary only along a single dimension, more complex contrasting cases can include several variations, thus exposing students to trade-offs and more complex schemas.

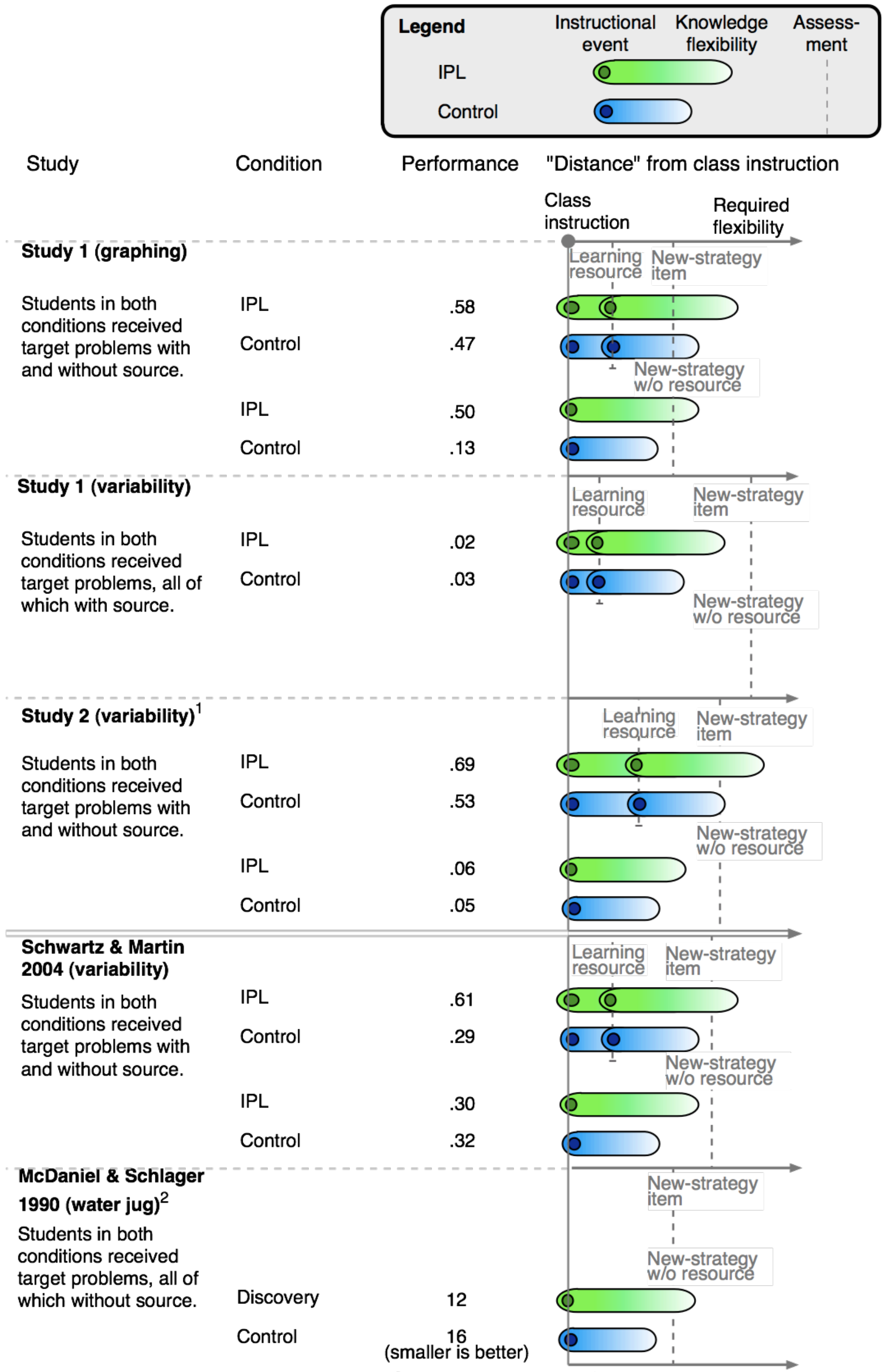
Another important factor is classroom culture. It seems that a forgiving, creative classroom culture is essential for the success of IPL. It seems that IPL both benefits from and contributes to such environment. This culture may explain the motivational effect of IPL on students with high-test anxiety in Study 1.

Study 1 provided evidence that simply analyzing the contrasting cases does not yield the cognitive or motivational benefits of IPL. Study 2 examined the role of design in the process. While the results of study 2 are inconclusive (probably due to lack of power), their patterns are identical to the findings in Study 1, suggesting that the design phase is a crucial component of IPL. In other words, it seems that students benefit from IPL as long as they get to design their own methods and evaluate these. Benefits of generation were found elsewhere in the memory and motivation literature, and usually have a procedural account (McNamara & Healy, 2000; Richland, Bjork, Finley & Linn, 2005). However, this thesis joins few other lines of research in extending the importance of generation to more conceptual tasks (c.f., Hausmann & VanLehn, 2007). It is quite likely that a combination of the conditions in study 2 would yield best results - that is, to let students design novel methods based on pre-designed faulty ones. This may save time, as well as direct students' attention to relevant procedural components.

6.2 So what IS the effect of IPL?

Both studies found a similar pattern of results, in which advanced students appeared to improve their ability to solve new-strategy items and debug faulty procedures, but with no effect on their performance on isomorphic and near transfer problems. Results in the regular classes lacked such a clear pattern. In some cases, IPL students performed better on isomorphic measures, compared with control (first unit in Study 1 and in the tutor unit in Study 2). In another case the opposite was true (second unit in Study 1). It seems that IPL has a strong dependency on prior knowledge of several sorts. First, students should have sufficient mathematical proficiency to invent methods. This includes understanding of general mathematical structures, as well as domain specific building blocks. Second, students should have an intuitive understanding of the target concept and the task at hand (i.e., be able to rank the cases accurately and reasonably consistently). Last, students should have at least some understanding of relevant aspects of the scientific method. The role of the latter is to help students link their mathematical knowledge with their qualitative understanding, intuition and experiences.

When these conditions are met, IPL can have a positive effect on students' knowledge. It appears that IPL does not simply help students gain "more of the same". Instead, it seems to modify the type of knowledge students acquire. More specifically, it helps students acquire more flexible knowledge that can be adapted and expanded as needed. In other words, IPL is designed to help students learn in situations that are further away from their prior learning events. This may help students better encode new instruction as seen in Study 2 and in Schwartz and Martin (2004), or it can help students spontaneously solve tasks that require novel strategies (as seen in Study 1 and in McDaniel and Schlager, (1990)). Figure 20 illustrates how this framework explains the different results found in these studies. The figure plots assessments in terms of ‘distance’ from original instruction, where ‘further’ assessments require higher flexibility of knowledge. The role of the embedded learning resource, according to this framework, is to mediate the gap between the new-strategy item and the preceding instruction. Figure 20 shows that the different results reported in study 1, study 2, Schwartz and Martin (2004), and McDaniel and Schlager (1990) can be explained using a unified framework that assumes that invention tasks contribute to more flexible knowledge. The consistency of the results between studies 1 and 2 (and their compatibility with prior research) suggests that IPL can indeed be systematically replicated to achieve comparable results.



Study	Condition	Performance
Study 1 (graphing)		
Students in both conditions received target problems with and without source.	IPL	.58
	Control	.47
	IPL	.50
	Control	.13
Study 1 (variability)		
Students in both conditions received target problems, all of which with source.	IPL	.02
	Control	.03
Study 2 (variability)[1]		
Students in both conditions received target problems with and without source.	IPL	.69
	Control	.53
	IPL	.06
	Control	.05
Schwartz & Martin 2004 (variability)		
Students in both conditions received target problems with and without source.	IPL	.61
	Control	.29
	IPL	.30
	Control	.32
McDaniel & Schlager 1990 (water jug)[2]		
Students in both conditions received target problems, all of which without source.	Discovery	12
	Control	16 (smaller is better)

1. Control is the average of the two no-design conditions.
2. Discovery group did not receive subsequent instruction.

Figure 20: Consolidating results from IPL studies. The diagram illustrates a possible explanation for the effect of IPL, namely, that IPL helps students acquire more flexible knowledge. As seen in the diagram, this explanation can help consolidate the results form the studies described in this thesis and previous research. Learning resources help students bridge their existing knowledge with the knowledge required for the new-strategy items.

The IPL process seems to have a positive effect also on students' motivation to learn. Study 1 found that students in the Full Invention condition remained working voluntarily during break time. This was true especially in between sessions of invention, and even though students were not graded for their inventions. This suggests not only more motivation, but also the adoption of more learning oriented goals (as suggested by Belenky and Nokes, 2009). In study 2 about 20% of the no-design students were observed designing novel methods rather than evaluating pre-designed methods, thus suggesting that students preferred designing new methods to evaluating existing ones. This may be another sign for increased motivation, or simply be the result of the existing methods being on paper rather than embedded in the system. Study 1 also found self-report evidence for the motivational benefits of IPL, especially for students with high test-anxiety. While the results mentioned above suggest that IPL has motivational benefits, a more detailed account is required to explain these effects.

Both studies suggest that students did not acquire better invention or sense-making skills. The invention lab especially failed heroically to help students become better scientists. Though the invention process was scaffolded, students did not receive instruction on it, and it was not framed as a learning goal. In a different study, we found that students began to internalize a different metacognitive construct, help-seeking behavior, only after receiving explicit support in multiple domains (Roll et al., submitted). It may be that a sequence of invention tasks on multiple topics will help students become better scientists. An open question is whether meta-IPL should be applied to the IPL process itself - that is, whether students should attempt to invent the scientific method before being taught it.

6.3 Explaining the effect of IPL

One of the main goals of this thesis is to identify the mechanism in which IPL achieves its effects. Table 25 summarizes the support that studies 1 and 2 found for the different hypothesis.

Table 25: Evaluating the predictions based on studies 1 & 2

Hypothesis	Predictions	Evidence in study 1	Evidence in study 2
H1: Self regulated learning hypothesis	• IPL students are more likely to attempt new challenges.	X	(not assessed)
	• IPL students perform better on invention tasks in a different domain.	(not assessed)	X
H2: Motivation hypothesis	• IPL students are more motivated to learn (and are especially more likely to adopt mastery goals)	✓	✓-
	• There is a significant correlation between motivational measures and learning outcomes.	X	(not assessed)
H3: Domain knowledge hypothesis	• There is direct mapping between features identified by students during invention attempts and features required by assessment items that evaluate flexible knowledge.	✓	(not assessed)
	• IPL students are more capable of diagnosing errors in variations on procedures learned in class.	✓	✓
H4: Impasse hypothesis	• Students who reach an impasse during invention perform better during assessment.	X	X
	• Reaching an impasse has the largest effect on knowledge that directly resolves the impasse.	X	X

The first hypothesis, H1, suggested that IPL students acquire better invention skills, and thus are better posed to invent during the test. However, as noted above, this is probably not the case. Study 1 shows that IPL students are not more likely to attempt new-strategy items, as would be expected based on this hypotheses. Study 2 showed that there were no fundamental differences between conditions with regard to invention behavior within the Invention Lab or during an invention task in a new domain during the test.

The second hypothesis, H2, suggested that a combination of domain-level gains helps IPL students perform better on debugging and future learning assessments. At the procedural level, experimenting with the different methods help students acquire better understanding of the functions that procedural components fill. At the conceptual level, students encounter more features during invention, and integrate them better thanks to the mathematical formality. This helps them acquire more elaborated schemas during class instruction.

The results from both studies support this hypothesis on multiple accounts. Improved performance on debugging items applies that IPL students indeed have superior functional understanding of the procedures. Better performance on new-strategy items suggests that IPL are better at integrating the new knowledge to their existing schemas, and adapting existing procedures to the new challenges. Additional support to this hypothesis comes from mapping of features that were encountered during inventions to features that were used during new-strategy assessments. Overall, it seems that though students fail to invent valid methods, they acquire key procedural and conceptual knowledge components, that help them extend the flexibility of their knowledge.

The third hypothesis, H3, suggested that a motivational explanation to IPL. This explanation argues that IPL leads to improved self-efficacy and adoption of mastery goals. Both studies found motivational benefits for IPL. Behavior during study 1 is associated with mastery goals, and the self reports found an improvement in students' situational interest. However, study 2 failed to find motivational benefits according to students' self reports, and its effect on students' goal orientation is not clear. Furthermore, this hypothesis does not make detailed predictions about the type of tasks in which IPL students would show improvement. Last, study 1 did not find correlation between the different motivational measures and performance on new-strategy items. The motivational benefits of IPL are important in and of themselves, and may contribute to part of the cognitive success of IPL. Yet, it is unlikely that motivation is the sole factor behind the cognitive effect of IPL. Cognitive changes are needed to yield differential performance improvements.

A forth hypothesis, H4, suggested that students reach impasses during inventions, and thus are more prepared to learn from the subsequent direct instruction. Study 1 provided evidence against this hypothesis. First, students in both conditions in study 1 had the opportunity to reach impasses. During the intuitive ranking stage, students in both conditions attempted to apply their previous knowledge, and in both topics (graphing and variability) shallow heuristics failed to distinguish between the contrasting cases appropriately (e.g., range). This was also emphasized during the subsequent class discussion, which took place in both conditions. According to the impasse hypothesis, students in both conditions should have reached an impasse, and thus no differences in learning would have been expected (on the contrary, Ranking Only students were supposed to learn more, given their longer show-and-practice instruction). However, the study results found the opposite, namely, that Ranking Only students did not learn as well as Full Invention students. The other piece of evidence that weakens the impasse hypothesis is the pattern of results. Based on the impasse hypothesis, one may predict that the effect of the impasse is the greatest on instruction that resolves the impasse. Therefore, the observable effect of IPL should be the greatest on isomorphic measures, since these are the closest to the impasse students reached. However, the trend in results was the opposite, with no consistent effect on isomorphic measures, and large effect on new-strategy items that do not share much in common with the impasse.

Overall, while students may indeed encounter productive impasses, this is probably not the cause for the cognitive effect of IPL.

6.4 Using technology to facilitate IPL

The work described in this thesis made headway in facilitating IPL using technology. It provides a proof of existence by building the Invention Lab and demonstrating its ability to direct the invention activity. The Invention Lab analyzes inventions and generates contrasting cases while supporting the exploratory nature of invention tasks. Students' inventions from the pilot and from Study 2 suggest that the lab enabled methods similar to the ones invented using paper and pencil in study 1. The progression in students' thinking as seen in Table 21 demonstrates the rich, diverse, and meaningful experiences students encounter while working with the lab. Still, the data from Study 2 also raises some concerns about the cognitive load imposed by the lab and its motivational effect. Lacking appropriate control condition, it is not clear whether the trend in results on new-strategy items is due to the Invention Lab or in spite of it. Longer studies with the lab will need to be conducted before these questions can be fully answered. However, the

Invention Lab in its current version already shows that an ITS can be used to facilitate the invention process while giving adaptive feedback and without handicapping its open-ended nature.

6.5 Generalizability of IPL

The set of studies described in this thesis provides strong evidence that IPL can be facilitated with different researchers and different populations. At the same time, these studies raise new questions. For example, the studies described above do not evaluate IPL in domains other than data analysis. Of main concern is the relative lack of improvement in the regular classes. IPL did not fail in these classes, since IPL students performed as well as (and at times better than) their peers despite shorter show-and-practice instruction. Still, the IPL trademark, that is, improved performance on new-strategy items, was not observed for regular students in neither study. One option is that the materials did not target that population well enough, and simpler tasks should have been given. An alternative explanation suggests that there is something qualitatively different that prevents IPL from succeeding with low-achieving students. The simple question that future studies will have to address is the following: what is the prior knowledge that is needed for successful IPL?

Chapter 7
Summary and Contributions

The IPL process offers an intriguing combination of inquiry tasks and direct instruction. Early studies with IPL instruction found improved performance on new-strategy tasks, an important outcome in today's dynamic world. The set of studies described above makes the first steps in unpacking IPL by identifying its core task elements and corresponding cognitive processes, and evaluating its scalability (across demographics, researchers, and mediums). So far, most evaluations of comparable constructivist manipulations were qualitative, lacked control, were done in a sterile setting, or were merely philosophical. This thesis is unique in applying analytic approach to studying the invention process, while keeping students in their natural environment - the classroom.

This thesis evaluates the effect of IPL and explains its causes, and thus belongs in the applied and basic research node of the Pasteur's Quadrant (Stokes, 1997). Due to its interdisciplinary nature, this thesis makes contributions to several related fields. In the field of cognitive science, the thesis improves our understanding of invention behavior and outcomes, and demonstrates a conceptual cognitive model of domain knowledge. In the field of the learning sciences, this thesis extends the empirically-based discussion around direct vs. constructivist instruction by supporting their combination rather than an either/or approach. This thesis also makes methodological contributions by developing assessments for flexibility of knowledge and invention skills, and by applying an analytic approach to a design problem. This work makes contributions to the fields of human-computer interaction and user modeling. Educational systems typically coach constrained problem solving or offer no adaptive support at all. This thesis introduces the Invention Lab, a one of its kind intelligent inquiry environment, and evaluates it in the field.

Overall, this thesis demonstrates the large potential of IPL to systematically improve students' flexibility of knowledge and motivation towards learning. It also explains this effect, and it identifies an opportunity and ability to facilitate IPL using an intelligent tutoring system. At the same time, the thesis points out many challenges. For example, it does not solve the problem of applying IPL in low-achieving classrooms. Also, and of main interest to me, are questions regarding the invention process itself. While IPL lays the framework for using the scientific method in the classroom, students do not yet capitalize on this opportunity. The challenge of making students better scientists remains largely unmet.

Chapter 8

References

1. Aleven, V., & Ashley, K. D. (1997). Teaching case-based argumentation through a model and examples empirical evaluation of an intelligent learning environment. In Artificial intelligence in education, 1997: Knowledge and media in learning systems: Proceedings of AI-ED 97, world conference on artificial intelligence in education, Kobe, Japan.
2. Aleven, V., & Koedinger, K. R. (2000). Limitations of student control: Do students know when they need help? In 5Th international conference on intelligent tutoring systems. Berlin: Springer Verlag.
3. Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. Cognitive Science, 26(2), 147-179.
4. Aleven, V., McLaren, B. M., Sewall, J., & Koedinger, K. R. (2006). The cognitive tutor authoring tools (CTAT): Preliminary evaluation of efficiency gains. In proceedings of the International conference on intelligent tutoring systems. Berlin: Springer Verlag.
5. Aleven, V., McLaren, B. M., Sewall, J., & Koedinger, K. R. (2008). Example-Tracing tutors: A new paradigm for intelligent tutoring systems. International Journal of Artificial Intelligence in Education.
6. Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. Journal of the Learning Sciences, 4(2), 167-207.
7. Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. R. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. Journal of Interactive Learning Research, 19(2), 185-224.
8. Barab (2005). Design-Based research: A methodological toolkit for the learning scientist. In Cambridge handbook of the learning sciences. (pp. 325-57).
9. Bargh, J. A., & Schul, Y. (1980). On the cognitive benefits of teaching. Journal of Educational Psychology, 72(5), 593-604.
10. Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. Psychological Bulletin, 128(4), 612-637.
11. Belenky, D. M., & Nokes, T. J. (2009). Motivation and transfer: The role of achievement goals in preparation for future learning. In Proceedings of the 31st Annual Conference of the Cognitive Science Society. (pp. 1163-8). Cognitive Science Society.
12. Bransford, J. D., & Schwartz, D. L. (2001). Rethinking transfer: A simple proposal with multiple implications. Review of Research in Education, 24(3), 61-100.
13. Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). Learning and transfer. In How people learn: Brain, mind, experience, and school. (pp. 51-78). NAP.
14. Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. The Journal of Learning Sciences.
15. Butcher, K. R., & Aleven, V. (2008). Diagram interaction during intelligent tutoring in geometry: Support for knowledge retention and deep understanding. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), Proceedings of the 30th annual conference of the cognitive science society. (pp. 1736-41). Austin, TX: Cognitive Science Society.
16. Chen, Z., & Klahr, D. (2008). Remote transfer of scientific-reasoning and problem-solving strategies in children. Advances in Child Development and Behavior, 36, 419-70.
17. Chi, M. T. H., De Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. Cognitive Science, 18(3), 439-477.
18. Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121-152.
19. Collins, A., & Halverson, R. (2009). Rethinking education in the age of technology: The digital revolution and schooling in America. New York, NY: Teachers College Press.
20. Collins, A., Brown, J. S., & Holum, A. (1991). Cognitive apprenticeship: Making thinking visible. American Educator, 15(3), 6-11.

21. Collins, A., Joseph, D., & Bielaczyc, K. (2003). Design research: Theoretical and methodological issues. Design-Based Research: Clarifying the Terms, 15.
22. Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4(4), 253-278.
23. Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. Journal of Educational Psychology, 88(4), 715-730.
24. Dewey, J. (1964). John Dewey on education: Selected writings. New York, NY: Modern Library.
25. Elliot, A. J., McGregor, H. A., & Gable, S. (1999). Achievement goals, study strategies, and exam performance: A mediational analysis. Journal of Educational Psychology, 91(3), 549-563.
26. Friedman-Hill, E. (2003). Jess in action: Rule-Based systems in java. Greenwich, CT: Manning.
27. Gama, C. (2004). Metacognition in interactive learning environments: The reflection assistant model. In proceedings of the International conference on intelligent tutoring systems. Berlin Heidelberg: Springer-Verlag.
28. Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. Journal of Educational Psychology, 95(2), 393-408.
29. Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning; differentiation or enrichment?. Psychological Review, 62(1), 32-41.
30. Halpern, D. F. (1998). Teaching critical thinking for transfer across domains. American Psychologist, 53(4), 449-455.
31. Hatano, G., & Inagaki, K. (1986). Two courses of expertise. In H. Stevenson, H. Azuma, & K. Hakuta (Eds.), Child development and education in Japan. (pp. 262-72). NY: Freeman.
32. Hausmann, R. G. M., & VanLehn, K. (2007). Explaining self-explaining: A contrast between content and generation. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), Artificial intelligence in education: Building technology rich learning contexts that work. Amsterdam, The Netherlands: IOS Press.
33. Heckler, A. F., Kaminski, J. A., & Sloutsky, V. M. (2008). Learning associations that run counter to biases in learning: Overcoming overshadowing and learned inattention. In Proceedings of the XXX annual conference of the cognitive science society. (pp. 511-6). Austin, TX: Cognitive Science Society.
34. Heffernan, N. T., & Koedinger, K. R. (1997). The composition effect in symbolizing: The role of symbol production vs. Text comprehension. In Proceedings of the Nineteenth annual conference of the cognitive science society. Hillsdale, NJ: Erlbaum.
35. Heffernan, N. T., & Koedinger, K. R. (1998). A developmental model for algebra symbolization: The results of a difficulty factors assessment. In Proceedings of the twentieth annual conference of the cognitive science society.
36. Hiebert, J., Stigler, J. W., Jacobs, J. K., Givvin, K. B., Garnier, H., Smith, M., et al. (2005). Mathematics teaching in the United States today (and tomorrow): Results from the TIMSS 1999 video study. Educational Evaluation and Policy Analysis, 27(2), 111-132.
37. de Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. Review of Educational Research, 68, 179-201.
38. Jean Piaget. (2009, October 20). In Wikipedia, the free encyclopedia. Retrieved October 21, 2090, from http://en.wikipedia.org/wiki/Jean_Piaget
39. Judd, C. H. (1908). The relation of special training to general intelligence. Educational Review, 36, 28-42.
40. Kapur, M. (2008). Productive failure. Cognition and Instruction, 26(3), 379 - 424.
41. Kapur, M., & Lee, K. (2009). Designing for productive failure in mathematical problem solving. In Proceedings of the 31st annual conference of the cognitive science society. (pp. 2632-7). Austin, TX: Cognitive Science Society.
42. Kieras, D. E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. Cognitive Science, 8(3), 255-73.

43. Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational Psychologist, 41(2), 75-86.
44. Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. Cognitive Science, 12(1), 1-48.
45. Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction - effects of direct instruction and discovery learning. Psychological Science, 15(10), 661-667.
46. Koedinger, K. R. (2002). Toward evidence for instructional design principles: Examples from cognitive tutor math 6. In Invited paper in PME-NA XXXIII (the north American chapter of the international group for the psychology of mathematics education).
47. Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. Educational Psychology Review, 19(3), 239-264.
48. Koedinger, K. R., & VanLehn, K. (2006). Pittsburgh science of learning center strategic plan.
49. Koedinger, K. R., Aleven, V., Roll, I., & Baker, R. S. J. d. (In press). In vivo experiments on whether supporting metacognition in intelligent tutoring systems yields robust learning. Handbook of Metacognition in Education.
50. Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. International Journal of Artificial Intelligence in Education, 8, 30-43.
51. Koedinger, K. R., Suthers, D. D., & Forbus, K. D. (1999). Component-Based construction of a science learning space. International Journal of Artificial Intelligence in Education, 10, 292-313.
52. Kuhn, D. (2007). Is direct instruction the answer to the right question?. Educational Psychologist, 42(2), 109-113.
53. Kuhn, D., & Pearsall, S. (2000). Developmental origins of scientific thinking. Journal of Cognition and Development, 1(1), 113-129.
54. Leelawong, K., & Biswas, G. (2008). Designing learning by teaching agents: The betty's brain system. International Journal of Artificial Intelligence in Education, 18(3), 181-208.
55. Lesgold, A., aajoie, S., Bunzo, M., & Eggan, G. (1991). SHERLOCK: A coached practice environment for an electronics troubleshooting job. In J. H. :arkin, & R. W. Chabay (Eds.), Computer assisted instruction and intelligent tutoring systems: Shared goals and complementary approaches. Hillsdale, NJ: Lawrence Erlbaum Associates
56. Luckin, R., & du Boulay, B. (1999). Ecolab: The development and evaluation of a vygotskian design framework. International Journal of Artificial Intelligence in Education, 10(2), 198-220.
57. Mathan, S. A., & Koedinger, K. R. (2005). Fostering the intelligent novice: Learning from errors with metacognitive tutoring. Educational Psychologist, 40(4), 257-265.
58. Matlen, B. J., & Klahr, D. (2009). Effects of instructional sequence on children's acquisition of the control of variables strategy. Carnegie symposium festschrift for David Klahr. Pittsburgh, PA.
59. McDaniel, M. A., & Schlager, M. S. (1990). Discovery learning and transfer of problem-solving skills. Cognition and Instruction, 7(2), 129-159.
60. McNamara, D. S., & Healy, a. f. (2000). A procedural explanation of the generation effect for simple and difficult multiplication problems and answers. Journal of Memory and Language, 43, 652-679.
61. Minstrell (2001). The role of the teacher in making sense of classroom experiences and effecting better learning. In Cognition and instruction, 25 years of progress. (pp. 121-49). Mahwah, NJ: Erlbaum.
62. Mitchell, M. (1992). Situational interest: Its multifaceted structure in the secondary mathematics classroom. Journal of Educational Psychology, 85(3), 424-436.
63. Mitrovic, A., & Ohlsson, S. (1999). Evaluation of a constraint-based tutor for a database language. International Journal of Artificial Intelligence in Education, 10(3-4), 238-256.

64. Morgan, P., & Ritter, S. (2002). An experimental study of the effects of cognitive tutor algebra I on student knowledge and attitude. Pittsburgh, PA: Carnegie Learning Inc. Retrieved December 2, 2008, from http://www.carnegielearning.com/wwc/originalstudy.pdf
65. Mott, B. W., & Lester, J. C. (2006). Narrative-Centered tutorial planning for inquiry-based learning environments. LECTURE NOTES IN COMPUTER SCIENCE, 4053, 675.
66. Nathan, M. J. (1998). Knowledge and situational feedback in a learning environment for algebra story problem solving. Interactive Learning Environments, 5(1), 135-159.
67. Nathan, M. J., Kintsch, W., & Young, E. (1992). A theory of algebra-word-problem comprehension and its implications for the design of learning environments. Cognition and Instruction, 9(4), 329-389.
68. Nussbaum, J., & Novick, S. (1982). Alternative frameworks, conceptual conflict and accommodation: Toward a principled teaching strategy. Instructional Science, 11(3), 183-200.
69. Ohlsson, S. (1994). Constraint-Based student modeling1. Student Modeling: The Key to Individualized Knowledge-Based Instruction, 167.
70. Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. Cognition and Instruction, 1(2), 117-175.
71. Papert, S. (1980). Mindstorms: Children, computers, and powerful ideas. Basic Books, Inc. New York, NY, USA.
72. Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. Educational Psychologist, 36(2), 89-101.
73. Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. Journal of Educational Psychology, 82(1), 33-40.
74. Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). Educational and Psychological Measurement, 53(3), 801-813.
75. Popper, K. R. (2002). Conjectures and refutations : The growth of scientific knowledge. London ; New York : Routledge. (Original work published 1963)
76. Reif, F., & Scott, L. A. (1999). Teaching scientific thinking skills: Students and computers coaching each other. American Journal of Physics, 67(9), 819-831.
77. Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). Linking cognitive science to education: Generation and interleaving effects. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), Twenty-Seventh annual conference of the cognitive science society. Mahwah, NJ: Lawrence Erlbaum.
78. Rittle-Johnson, B. (2004). Promoting flexible problem solving: The effects of direct instruction and self-explaining. In Proceedings of the 26th annual meeting of the cognitive science society. Mahwah, NJ: Erlbaum.
79. Rittle-Johnson, B., & Star, J. R. (2009). Compared with what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving. Journal of Educational Psychology, 101(3), 529-544.
80. Roll, I., Aleven, V., & Koedinger, K. R. (2009). Helping students know 'further' - increasing the flexibility of students' knowledge using symbolic invention tasks. In N. A. Taatgen, & H. van Rijn (Eds.), Proceedings of the 31st annual conference of the cognitive science society. (pp. 1169-74). Austin, TX: Cognitive Science Society.
81. Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2007). Designing for metacognition - applying cognitive tutor principles to the tutoring of help seeking. Metacognition and Learning, 2(2), 125-140.
82. Salden, R. J. C. M., Aleven, V. A., Renkl, A., & Schwonke, R. (2008). Worked examples and tutored problem solving: Redundant or synergistic forms of support? In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), Proceedings of the 30th annual conference of the cognitive science society. Austin, TX: Cognitive Science Society.
83. Savery, J. R., & Duffy, T. M. (1995). Problem based learning: An instructional model and its constructivist framework. In B. Wilson (Ed.), Constructivist learning environments: Case

studies in instructional design. (pp. 135-48). Englewood Cliffs, NJ: Educational Technology Publications.

84. Scardamalia, M., & Bereiter, C. (1994). Computer support for knowledge-building communities. The Journal of the Learning Sciences, 3(3), 265-283.
85. Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. Grouws (Ed.), Handbook of research on mathematics teaching and learning. (pp. 334-70). New-York: MacMillan.
86. Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2008). Motivation in education : Theory, research, and applications. Upper Saddle River, N.J. : Pearson/Merrill Prentice Hall.
87. Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. Cognition and Instruction, 16(4), 475-522.
88. Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. Cognition and Instruction, 22(2), 129-184.
89. Schwartz, D. L., Bransford, J. D., & Sears, D. A. (2005). Efficiency and innovation in transfer. In Mestre (Ed.), Transfer of learning from a modern multidisciplinary perspective. (pp. 1-51). CT: Information Age Publishing.
90. Schwartz, D. L., Martin, T., & Pfaffman, J. (2005). How mathematics propels the development of physical knowledge. Journal of Cognition and Development, 6(1), 65-88.
91. Schwartz, D. L., Sears, D., & Chang, J. (2007). Reconsidering prior knowledge. In M. C. Lovett, & P. Shah (Eds.), Thinking with data. (pp. 319-44). New York, NY: Routledge.
92. Scott, P. H., Asoko, H. M., & Driver, R. H. (1991). Teaching for conceptual change: A review of strategies. In R. Duit, F. Goldberg, & H. Neidderer (Eds.), Research in physics learning: Theoretical issues and empirical studies. (pp. 310-29). Keil, Germany: Schmidt & Klannig.
93. Sears, D. A. (2006). Effects of innovation versus efficiency tasks on recall and transfer in individual and collaborative learning contexts. In Int C on learning sciences.
94. Shute, V. J., & Glaser, R. (1990). A large-scale evaluation of an intelligent discovery world: Smithtown. Interactive Learning Environments, 1(1), 51-77.
95. Siegler, R. S. (1983). How knowledge influences learning. American Scientist, 71, 631-638.
96. Siegler, R. S. (2002). Microgenetic studies of self-explanation. In N. Granott, & J. Parziale (Eds.), Microdevelopment - transition processes in development and learning. Cambridge University Press.
97. Singley, M. K., & Anderson, J. R. (1989). Transfer in the ACT* theory. In The transfer of cognitive skills.
98. Star, J. R., & Rittle-Johnson, B. (2009). It pays to compare: An experimental study on computational estimation. Journal of Experimental Child Psychology, 102(4), 408-26.
99. Stokes, D. E. (1997). Pasteur's quadrant: Basic science and technological innovation. Brookings Institution Press.
100. Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. Educational Psychology Review, 10(3), 251-296.
101. Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. Psychological Review, (8), 247-261.
102. Tobias, S., & Duffy, T. M. (2009). Constructivist instruction: Success or failure? New York: Taylor & Francis.
103. Trumpower, D. L., & Fellus, O. (2008). Naive statistics: Intuitive analysis of variance. In Proceedings of cognitive science. (pp. 499-503).
104. van Joolingen, W. R. (1999). Cognitive tools for discovery learning. International Journal of Artificial Intelligence in Education, 10, 385-397.
105. VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., & Shelby, R. (2005). The Andes physics tutoring system: Five years of evaluation. In International conference on artificial intelligence in education.
106. VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring?. Cognition and Instruction, 2(3),

209-49.

107.Veermans, K., de Jong, T., & van Joolingen, W. R. (2000). Promoting self-directed learning in simulation-based discovery learning environments through intelligent support. Interactive Learning Environments, 8(3), 229-255.

108.Whewell, W., & Butts, R. E. (1989). Theory of scientific method. Indianapolis : Hackett Publication Company

109.White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. Cognition and Instruction, 16(1), 3-118.

110.White, B. Y., Shimoda, T. A., & Frederiksen, J. R. (1999). Enabling students to construct theories of collaborative inquiry and reflective learning: Computer support for metacognitive development. International Journal of Artificial Intelligence in Education, (10), 151-182.

111.Woolf, B. P., Marshall, D., Mattingly, M., Lewis, J., Wright, S., Jellison, M., et al. (2003). Tracking student propositions in an inquiry system. Artificial Intelligence in Education: Shaping the Future of Learning Through Intelligent Technologies, 21.

Chapter 9

Appendices

9.1 Features targeted by the Invention Lab

The cognitive model of the Invention Lab captures the following conceptual errors:

Target feature	Common errors	Algorithm for generating contrasting cases
The method should use all values	Extreme values: Method uses only extreme values	1. Keep losing case 2. Copy extreme values to other set as seed 3. Shift other points towards the mean: x -> (x+1.5mean)/2.5
	Sub range: Method uses only a sub range of the data	1. Keep used values as seed 2. Fill in up to a total of 5 values from regions above or below the sub-range, keeping the relative position of the sub-range.
	Min / max only: model only uses min or max.	Use algorithm for extreme values.
	Largest gap: the model uses only the subsequent numbers with the largest gap.	1. Keep largest gap as seed 2. Set 1: seed; max(seed)+2 units; max(seed) + 4 units; min(seed) - 3 units. 3. Set 2: seed; max(seed)+1 units; max(seed) + 2 units; min(seed) - 1 units.
	Other not all points	1. Keep used points as seed. 2. Set 1: seed; average seed + 1 (rounded); average seed - 1 (rounded). 3. Set 2: seed; halfway between max(seed) and 14 units; halfway between min(seed) and 1 unit.
Repeated values should be counted as the number of the repetitions	Ignore repeated values: each value is counted once by the model regardless of its repetitions	1. Seed = 3 values from the winner set: max(winner), min(winner), random (winner). Set 1: min, random, random, random, max. Set 2: min, min, random, max, max
	Gap count: the method counts the gaps between the numbers, and thus ignores repeating values	Same as above.

Variability is qualitatively different from central tendency	Students use only mean, median, mode.	1. Keep 1 set 2. Copy all values to set 2. 3. Shift set 2 up or down by 4 units
Method should control for number of points	Method does not control for number of points	Choose predesigned contrasts
Subtraction is a good way to evaluate distance	Method does not use subtraction	Choose predesigned contrasts
The method should use only values that appear in the dataset	Methods uses '0' to compensate for an odd number of numbers	1. Keep the set where '0' was used as seed 2. Copy all values to set 2 3. Add '0' to set 2.
	Method uses arbitrary constants	Same as above.

In addition, the cognitive model of the Invention Lab always checks the following:

- The same feature should not be targeted more than 3 times

- The difference between the MAD of the two cases should be at least 1.2 units

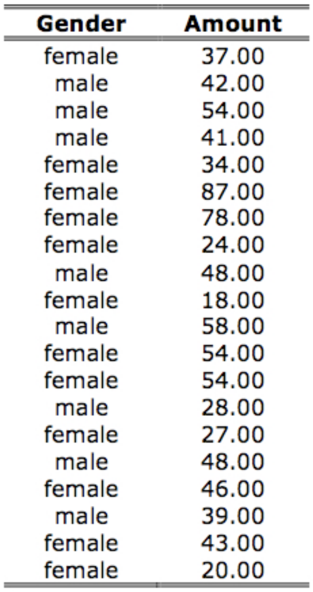
- The same case cannot be kept more than 3 times in a row.

9.2 Given tasks

This appendix includes the invention tasks that were given to students in both studies.

(a) Shop-O-Shirt (study 1, graphing & central tendency*)*

Ann and Dan sell clothes at the Shop-O-Shirt store at the Waterfront. They argue who spends more – boys or girls. Ann says that boys spend more. Dan thinks that girls spend more. Here is the list of the first 20 purchases people made at Shop-O-Shirt last Sunday:

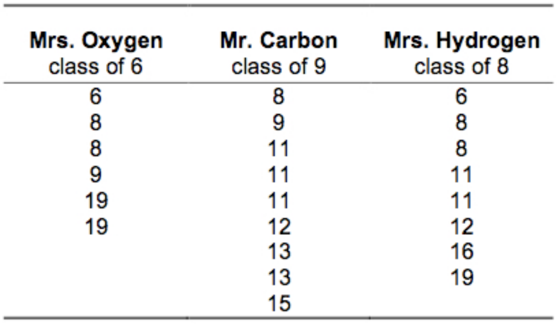


Gender	Amount
female	37.00
male	42.00
male	54.00
male	41.00
female	34.00
female	87.00
female	78.00
female	24.00
male	48.00
female	18.00
male	58.00
female	54.00
female	54.00
male	28.00
female	27.00
male	48.00
female	46.00
male	39.00
female	43.00
female	20.00

1. How much money they spend:
   a. Who is more likely to spend a lot of money at the Shop-O-Shirt? Boys / Girls
   b. Who is more likely to spend very little money at the Shop-O-Shirt? Boys / Girls
   c. Who is more likely to spend in the middle? Boys / Girls
2. Who spends more money at the Shop-O-Shirt per customer? Boys / Girls

(b) Making the Grade (study 1, graphing & central tendency)

Imagine your friend Devan is very worried about getting a good grade in Chemistry. He can take the class from Mrs. Oxygen, Mr. Carbon, or Mrs. Hydrogen. Here are the grades each teacher gave out last year on the final test (maximum was 20 points).



Mrs. Oxygen	**Mr. Carbon**	**Mrs. Hydrogen**
class of 6	class of 9	class of 8
6	8	6
8	9	8
8	11	8
9	11	11
19	11	11
19	12	12
	13	16
	13	19
	15	

1. What class do you think Devan should attend given that he is a good student?
__ - Mrs. Oxygen ___ - Mr. Carbon ___ - Mrs. Hydrogen

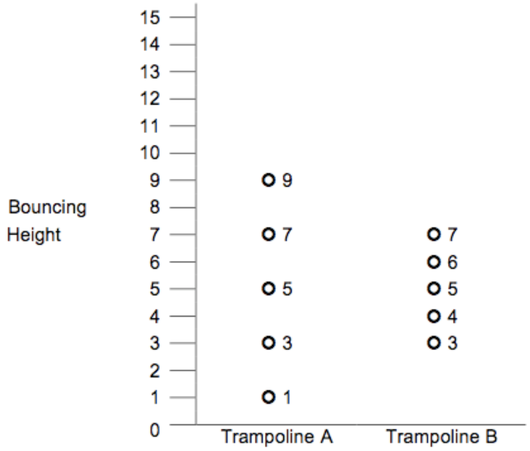
2. What class do you think Devan should attend given that he is only an okay student in chemistry?
__ - Mrs. Oxygen ___ - Mr. Carbon ___ - Mrs. Hydrogen

3. What class should Devan take if he is not sure how good he is in Chemistry? Which class is safer? Which class offers more opportunity?
The safer class:
__ - Mrs. Oxygen ___ - Mr. Carbon ___ - Mrs. Hydrogen
The class with more opportunity:
__ - Mrs. Oxygen ___ - Mr. Carbon ___ - Mrs. Hydrogen

The class Devan should choose if he is unsure of his ability:
__ - Mrs. Oxygen ___ - Mr. Carbon ___ - Mrs. Hydrogen

(c) The Olympic Trampoline. Study 1 and 2, variability

The Bouncers Trampoline Company makes several brands of trampoline. They test their trampolines by dropping a 100 lb weight from 15 feet. They measure how many feet the weight bounces back into the air. They do several trials for each trampoline.
Here are the results for two of their trampolines:



They like to report the results as sets:
Trampoline A: {1 3 5 7 9}
Trampoline B: {3 4 5 6 7}

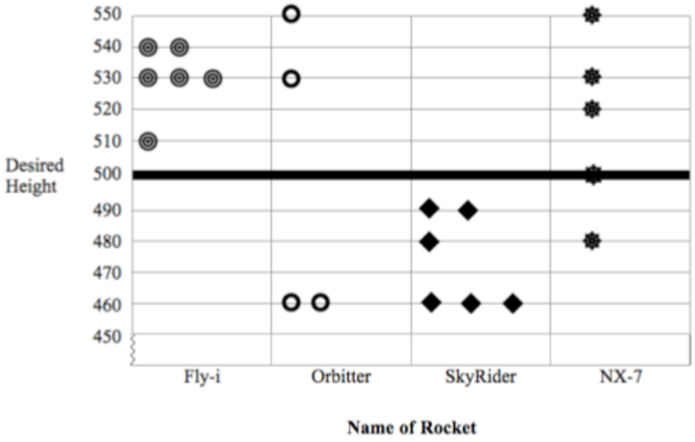
What is the average of Trampoline A? ___
What is the average of trampoline B? ___

The average is similar in both sets of numbers. But there is something different about these two sets of numbers. One group of numbers is more spread out. We say that the first set of numbers has a greater variance. The other group of numbers is closer together. We say that this trampoline has lower variance.
Which trampoline has a greater variance? A / B

- Study 1 also included the following contrasting cases:
  - {10, 2, 2, 10, 10, 2} vs. {2, 8, 4, 10, 6, 6}
  - {4, 2, 6} vs. {6, 2, 4, 2, 6, 4}
- Study 2 had dynamically-generated cases.

(d) NASA has a problem. Study 1, variability.

NASA needs to choose a rocket to launch its latest satellite. No rocket is currently ready – but NASA wants to choose one and focus on its development. At this point, NASA does not care about the absolute height it reaches, since the amount of fuel will need to be adjusted. NASA cares about the ability to predict what height the rocket will reach. They need a rocket that arrives at almost the same height every time.
The following graphs show the height the rockets reached during testing, relative to the desired height. Each line represents 10 miles. Each point represents the height a rocket reached in a single test.

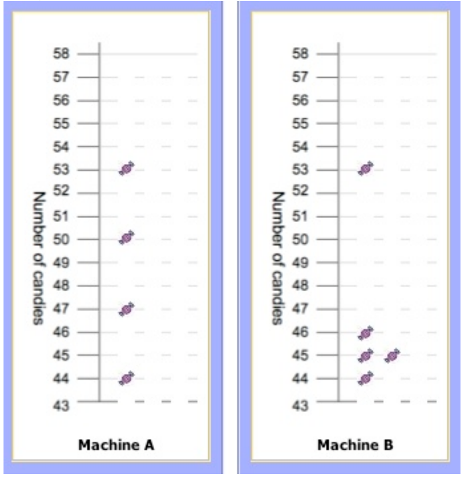


Which one do you think is the most appropriate missile for this task? (Circle one)

Fly-i / Orbitter / SkyRider / NX-7

(e) KanD. Study 2, variability.

A friend of mine has a small candy factory, called KanDee. Recently she needed a new machine that packs the candies in the packages. The problem: these machines are not accurate. Find a way to determine which machine is more accurate.



- Additional contrasting cases were generated dynamically.

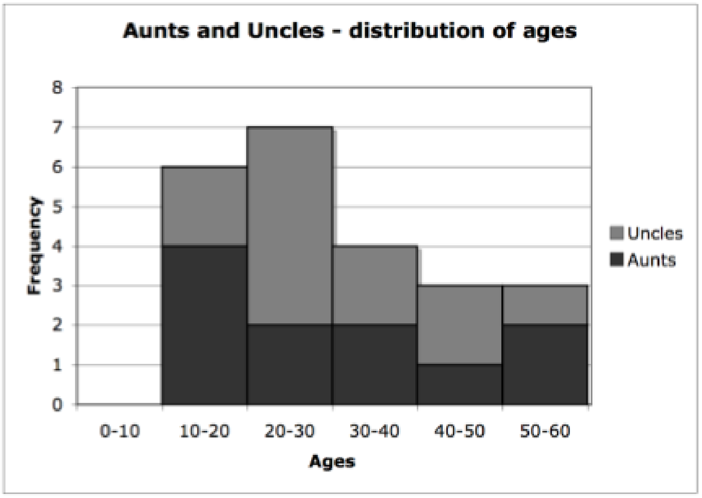
9.3 post-tests

(a) Study 1, topic 1 (graphing and central tendency)

(i) New strategy items

(1) Histograms - learning resource

Three friends made a histogram of the ages of their aunts and uncles. They picked different colors for their aunts and for their uncles:



The different color allows us to answer questions about each specific group or about the overall data.

For example:

- Question: How many aunts are between 30 and 40 years old?

Answer: 2 aunts. We look only at the darker gray that represents aunts.

Another example:

- Question: How many aunts and uncles altogether are between 20 and 30 years old?

Answer: 7. When the question asks about the overall of both groups, we look at the overall height of the histogram.

Please answer the following questions:

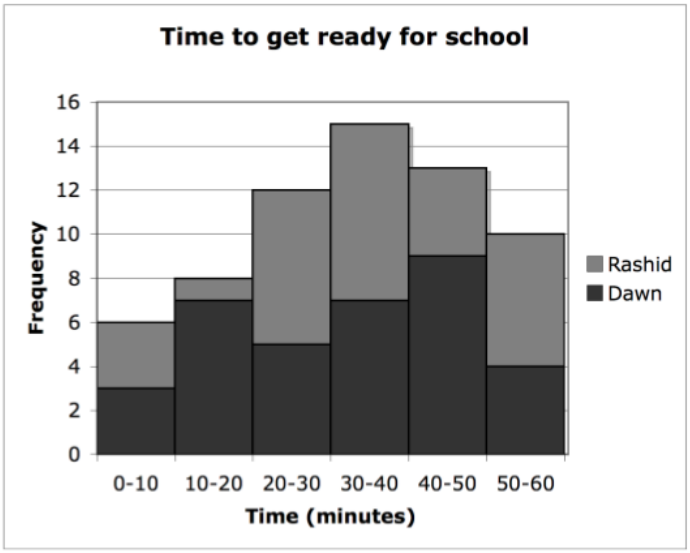
a. How many aunts and uncles altogether are between 40 and 50

b. How many aunts are between 10 and 20 years old?

c. How many uncles are between 10 and 20 years old?

(2) Histograms - new-strategy item

Dawn and Rashid asked their friends how long it takes them to get organized to school in a typical morning. Here are the answers they received:



a. How many of Dawn’s friends replied that they take less than 10 minutes to get organized?

b. What is the time range that seems most common overall for both Dawn’s and Rashid’s friends?
[ ] 30-40 minutes
[ ] 40-50 minutes
[ ] 50-60 minutes
[ ] The graph does not tell that.

c. What time frame is the LEAST typical amongst Rashid’s friends?

(3) Stem and leaf - learning resource

There are a couple of ways to apply stem-and-leaf plots to large numbers.
For example, how will you make a stem and leaf plot for the following set:
127 173 241 268 281 346

One option to put the hundreds and the tens in the Stem column, and to put the ones in Leaf column:

Stem	Leaf
10	
11	
12	7
13	
14	
15	
16	
17	3
18	
19	
20	
21	
...	

key: 12 | 7 = 127

However, this does not make much sense, since all we get is a very long list.
A better way for this set of numbers is to put the hundreds in the Stem column and the tens and ones in the Leaf column, as seen here:

Stem	Leaf
1	27 73
2	41 68 81
3	46

key: 1 | 27 = 127

Questions:
A. Make a stem-and-leaf plot for the following set (don’t forget to add the key!)

540 639 688 748 803 822 869

B. For each of the following cases, answer whether they describe the same data, and which stem-and-leaf plot is more informative:

Plot A:

Stem	Leaf
70	5
71	3 7
72	0 5 8
73	
74	6

key: 71 | 3 = 713

Plot B:

Stem	Leaf
7	05 13 17 20 25 28 46

key: 7 | 13 = 713

Do A and B represent the same data? Yes / No

If they represent the same data, which plot is more informative? A / B

Plot C:

Stem	Leaf
75	
76	3
77	
78	
79	1
80	
81	
82	9
83	
84	5
85	
86	
87	
88	6
89	
90	
91	
92	
93	1

key: 76 | 3 = 763

Plot D:

Stem	Leaf
7	63 91
8	29 45 86
9	31

key: 7 | 63 = 763

Do C and D represent the same data? Yes / No

If they represent the same data, which plot is more informative? C / D

(4) Stem and leaf - new-strategy item

The following table shows the NBA leaders for total points scored during playoffs, places 6-18.

Player name	Points
6 Larry Bird	3,897
7 John Havlicek	3,776
8 Hakeem Olajuwon	3,755
9 Magic Johnson	3,701
10 Scottie Pippen	3,642
11 Elgin Baylor	3,623
12 Wilt Chamberlain	3,607
13 Tim Duncan	3,282
14 Kevin McHale	3,182
15 Dennis Johnson	3,116
16 Julius Erving	3,088
17 Kobe Bryant	3,053
18 James Worthy	3,022

Make a stem-and-leaf plot of the data.

(ii) Other items

The Weather Report

a. What is the average of the temperatures in the following three cities in Israel?
Please show your work.

- Mount Hermon:	40 degrees Fahrenheit
- Tel Aviv:	70 degrees Fahrenheit
- Eilat:	130 degrees Fahrenheit

b. If we were to add Jerusalem to the cities, how would the average have changed?
The temperature in Jerusalem is between that of Mount Hermon and that of Tel Aviv.

The average would (please circle one):

[ ] Go up
[ ] Stay the same
[ ] Go down
[ ] I can't tell without knowing the exact number

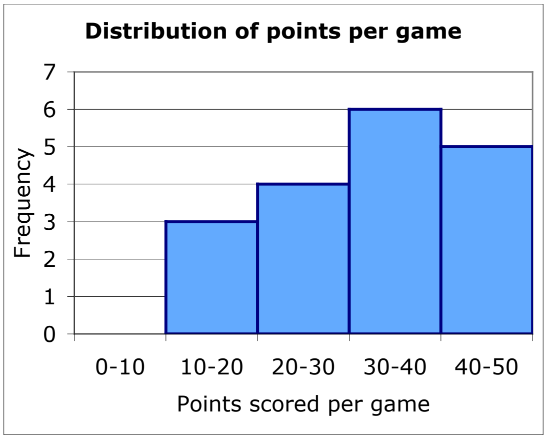
Temperatures at Jamestown

Tom wrote down the temperatures in Pine Hills during the month of October, so now he has a list organized in order of days (but not in order of temperatures). Which representation is most useful to answer each of the following questions?

A.	What was the temperature on March 2nd?	List / Box plot / Histogram
B.	What is the median of the temperatures?	List / Box plot / Histogram
C.	How many times was the temperature between 50 and 60°?	List / Box plot / Histogram

The Jaguars Basketball Team

Here is the distribution of points the Jaguars Basketball Team scored last year in their games:



a. How many games did they play overall?

b. In how many games did they score between 20 and 30 points?

c. What was more common – to score more than 40 point per game, between 30 and 40, or below 30?

[ ] Above 40 points per game
[ ] Between 30 and 40 points per game
[ ] Below 30 points per game
[ ] There is not enough information to answer that.

Matching graphs

A. Are the following statements correct? (More than one can be correct)

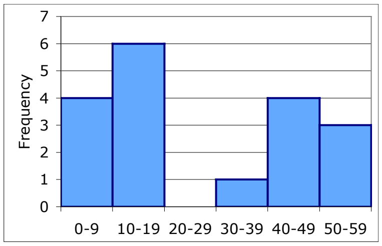
“The stem and leaf plot and Histogram A show the same data. “ True / False

“The stem and leaf plot and Histogram B show the same data. “ *True / False

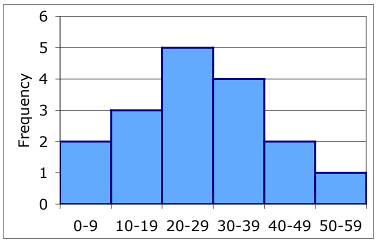
“The stem and leaf plot and Histogram C show the same data. “ *True / False

Stem	Leaf
0	3 8
1	5 8 9
2	1 2 4 4 6
3	2 6 7 8
4	2 5
5	7

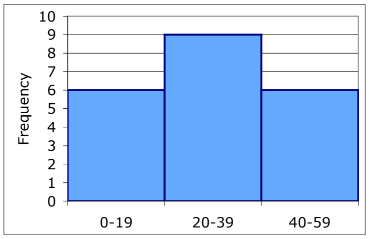
Histogram A:



Histogram B:



Histogram C:



B. Are the following statements correct? (More than one statement can be correct)

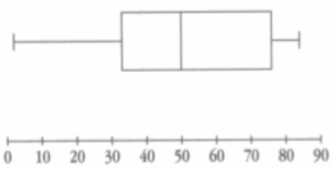
“The stem and leaf plot and box plot A show the same data.” True / *False

“The stem and leaf plot and box plot B show the same data.” *True / False

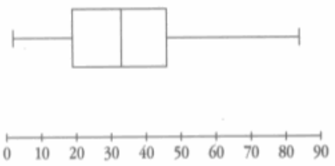
“The stem and leaf plot and box plot C show the same data.” True /* False

Stem	Leaf
0	2 3 3
1	4 5 6 7 7 9 9 9
2	2 2 3 5 5 5 8 8 9
3	2 4 4 5 6 6 6
4	0 0 1 2 6 7
5	1 2 3 3 4
6	0 5
7	1
8	4

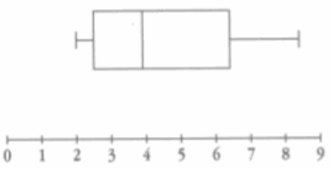
Box Plot A



Box Plot B



Box Plot C



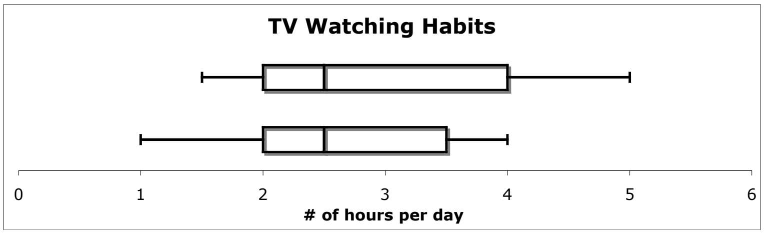
TV habits

[M]

During recess, the 7th grade students conducted a survey about TV watching habits. They asked the same number of students from 7th grade and from 6th grade the following question:

- *How many hours a day do you usually watch TV?*

In order to compare the two groups, they plotted the following box and whiskers graph.



a. Among the 7th graders, do more students watch less than 2 hours or more than 4 hours?
[ ] Less than 2 hours
[ ] More than 4 hours
[ ] About the same number of students
[ ] Cannot be told from the graph

b. In what group do more students watch between 2 and 4 hours of TV a day?
[ ] 7th grade students
[ ] 6th grade students
[ ] About the same number of students
[ ] Cannot be told from the graph

c. What is the **maximum** number of hours reported by 7th grade students?
[ ] 4
[ ] 5
[ ] 6
[ ] There is not sufficient data to answer the question.

b. What is the **median** number of hours reported by 7th grade students?
[ ] 2
[ ] 2.5
[ ] 3
[ ] There is not sufficient data to answer the question.

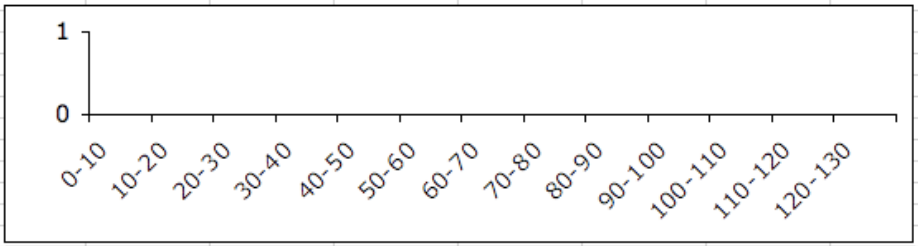
c. What is the **mean** number of hours reported by 7th grade students?
[ ] 2
[ ] 2.5
[ ] 3
[ ] There is not sufficient information to answer the question.

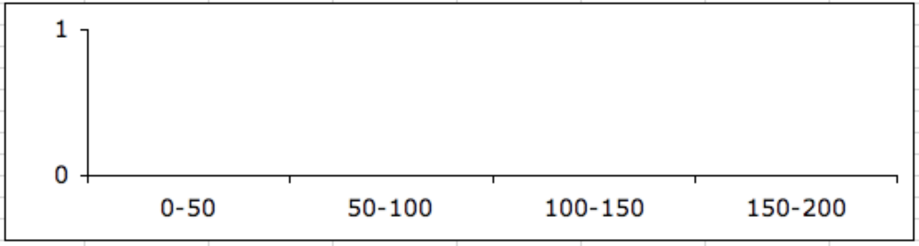
Choosing the scale

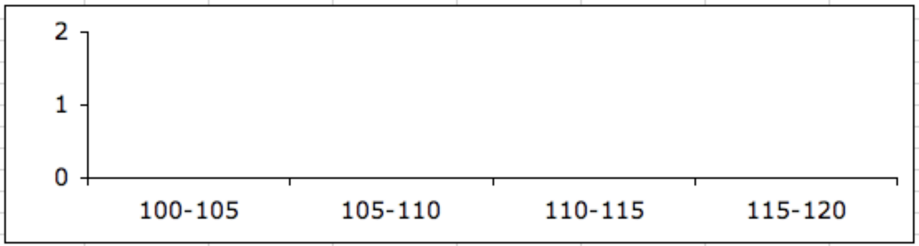
Jerry counted the number of peanuts in 12 packages of 1 Lb. He wanted to see the distribution and the typical range of number of peanuts in a package.
Here is what he got:

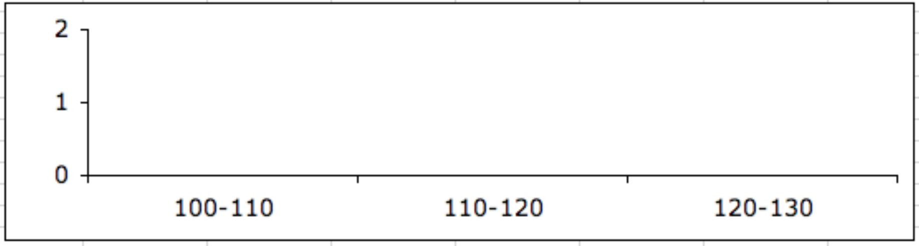
105 107 108 108 109 111 112 113 113 114 116 119

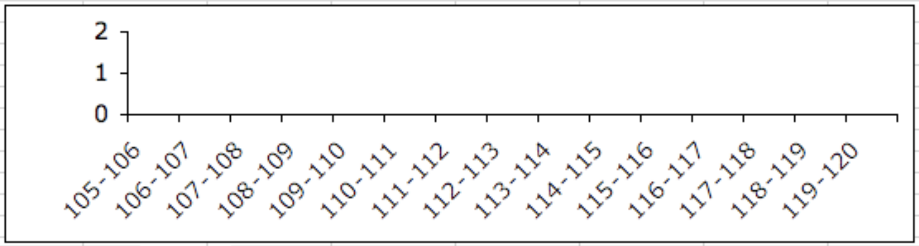
To answer his question, he wants to draw a histogram.
Which scale is most appropriate for this data? (circle one) A / B / C / D / E











Fighting the Flu

UPMC is testing a new medication to help reduce symptoms of cold and flu.
To do that, they first asked 11 people to answer the following question:

In how many days were you sick last month?

After collecting the responses, they gave all people the new medication.
After one month they asked the same people the same question again, to see how the medicine affected the number of days they were sick.

Here are the responses they got. Each number shows how many times that person was sick during that month.

Patient number:	1	2	3	4	5	6	7	8	9	10	11
Month 1 (before medicine):	13	7	9	15	21	7	16	15	14	24	13
Month 2 (after medicine):	21	23	21	18	5	8	22	6	14	3	8

Do you think that the medicine is worth taking?

(b) Study 1, topic 2 (variability)

(i) New strategy items

(1) Variability - learning resource

Sometimes MAD is not the best way to compare the variance of two sets. The MAD is not a good measure when one set has very large values while the other set has low values.
For example, look at the following two sets:

Set 1 (3 7 8)	Mean =6	MAD = 2
Set 2 (103 107 108)	Mean = 106	MAD = 2

The variance of Set 1 is much more important than the variance of Set 2.
In Set 1, the differences between the numbers are relatively big. In Set 2, the same differences are not as important, since they are small compared with the numbers.

MAD cannot tell this difference, since both sets have MAD = 2. We need a measure that describes the variance compared with the average.

When there are big differences between the averages of the sets, we use a measure that is called **Proportional MAD**
Proportional MAD is the MAD divided by the mean.

$$\text{Set 1:} \quad \text{Proportional MAD} = \frac{\text{MAD}}{\text{Mean}} = \frac{2}{6} = 0.3333$$

$$\text{Set 2:} \quad \text{Proportional MAD} = \frac{\text{MAD}}{\text{Mean}} = \frac{2}{106} = 0.0189$$

Since 0.333 is much bigger than 0.0189, the Proportional MAD of set 1 is much bigger than the Proportional MAD of set 2, which makes a lot of sense.

a. What is the **Proportional MAD** of set 3: (100 140 210)

b. When should you use Proportional MAD instead of regular MAD?

[ ] When the average of one set is very high and of the other set is very low

[ ] When the MAD is very high

[ ] When you cannot calculate the MAD

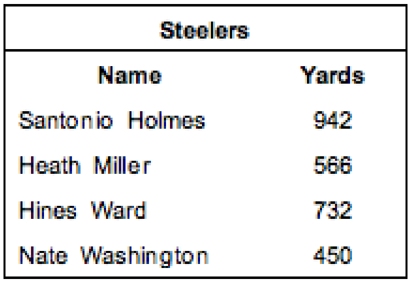
[ ] Never

(2) Variability - new-strategy item

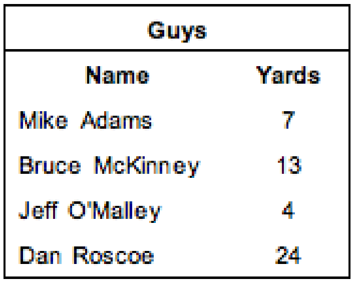
The Steelers vs. The Guys

Four friends, who often play football together (but are not very good at it), decided to compare their performance to the top Steelers players.

Here is the comparison of their overall receiving yards during the last season:



Steelers	
Name	**Yards**
Santonio Holmes	942
Heath Miller	566
Hines Ward	732
Nate Washington	450



Guys	
Name	**Yards**
Mike Adams	7
Bruce McKinney	13
Jeff O'Malley	4
Dan Roscoe	24

It is clear that the Steelers had many, many more yards. However, the guys were hoping that they at least had lower variability.

Given the huge differences in their overall receiving yards, who would you say has lower variance? Show your calculations.

(ii) Other items

Calculating MAD

Calculate the MAD of the following sets of numbers:
a. (5, 3, 8, 4, 5)

b. (25, 31, 34, 30)

Comparing prices

[C1]

Jamie wanted to compare prices in different locations of the same stores. To do this, she bought the same t-shirt at different Target and K-Mart stores.
Target:

- Number of Target stores she visited: 5
- Average price at Target: $7
- MAD at target: 4

K-Mart:

- Number of K-Mart stores she visited: 4
- Average price at K-Mart: $8.50
- MAD at K-Mart: 2

a. Which store has prices that are more similar between locations? Target / K-Mart
b. In which store do the prices have lower variance? Target / K-Mart
c. Which store is more expensive? Target / K-Mart

Spelling competition

[K]

Jerry participated in the school's spelling competition. To move to the next round, he needed an overall average of at least 10 points.
Jerry achieved the following scores:
(9, 16, 8)

a. Did he reach the required average? Does he qualify for the next round? *Yes / No

b. Jerry's math teacher asked his students to calculate the MAD of Jerry's scores.
For each student, write whether the method is correct or not.

Student A wrote the following:
Step 1: Jerry needs to reach an average of 10 points.
Step 2: I calculate the distances between the needed average and the scores he got. The distance between 9 and 10 = 1; between 16 and 10 = 6; between 8 and 10 = 2
Step 3: I do an average of these numbers. (1+6+2)/3 = 9/3 = 3. The MAD is 3.

Is this answer correct? Yes / No

If not, in what step did Student A make a mistake? Step 1 / 2 / 3

Student B wrote the following:
Step 1: I calculate the average Jerry got. (9+16+8)/3 = 33/3 = 11
Step 2: I subtract all the numbers from the average. 9-11 = (-2); 16-11=5; 8-11=(-3)
Step 3: I find the average of these numbers. ((-2)+5+(-3)) /3 = 0/3 = 0. The MAD is 0.

Is this answer correct? Yes / No

If not, in what step did Student A make a mistake? Step 1 / 2 / 3

Student C wrote the following:
Step 1: Jerry had an average of (9+16+8)/3 = 33/3 = 11
Step 2: The distance from the average to all the points is: Between 9 and 11 is 2; between 16 and 11 is 5; and between 8 and 11 is 3.
Step 3: I add up the three distances: (2+5+3)=10. The MAD is 10.

Is this answer correct? Yes / No

If not, in what step did Student A make a mistake? Step 1 / 2 / 3

Distance

[J]

Jason measured the distance from home to school in two methods.
Four times he counted steps, and got the following distances:
560, 520, 550 and 530 feet.
Three times he timed how long it takes him to run, and got the following distances:
470, 600, and 520 feet.

a. Which method was more reliable for Jason?

Running / Counting steps

Explain or show your calculations

b. Could Jason time his run only once and be sure he got the right distance?

Yes / No

Please explain.

What do you think the real distance is between Jason's home to school?

Order of heights]

Latoya, Alex, and Channel wanted to calculate the MAD of heights of students in their class. Each of them wrote down the heights of all students in their class, and then calculated the MAD. They followed the same steps, but had one little difference:

- Latoya organized the heights from the highest to the lowest.
- Alex organized the heights from the lowest to the highest
- Channel did not organize the heights in any special order

Which of the three students will have the highest MAD?

[ ] Latoya will get the highest MAD
[ ] Alex will get the highest MAD
[ ] Channel will get the highest MAD
[ ] They will all get the same MAD

Orange Juice Drinking Habits

The manager of the school cafeteria wants to study the drinking habits of the students.
He has three questions, and he knows how to calculate three different measures. But he needs some help using the right measure for the right question.
Please help him match the measure to the question

Questions: What measure? (circle one)

1. What is the typical quantity of OJ that one student drinks? MAD / Sum / Mean
2. How different are the quantities that the different students drink? MAD / Sum / Mean
3. How much overall OJ do the students drink altogether? MAD / Sum / Mean

Negative?

a. Can the average ever be negative? Yes / No

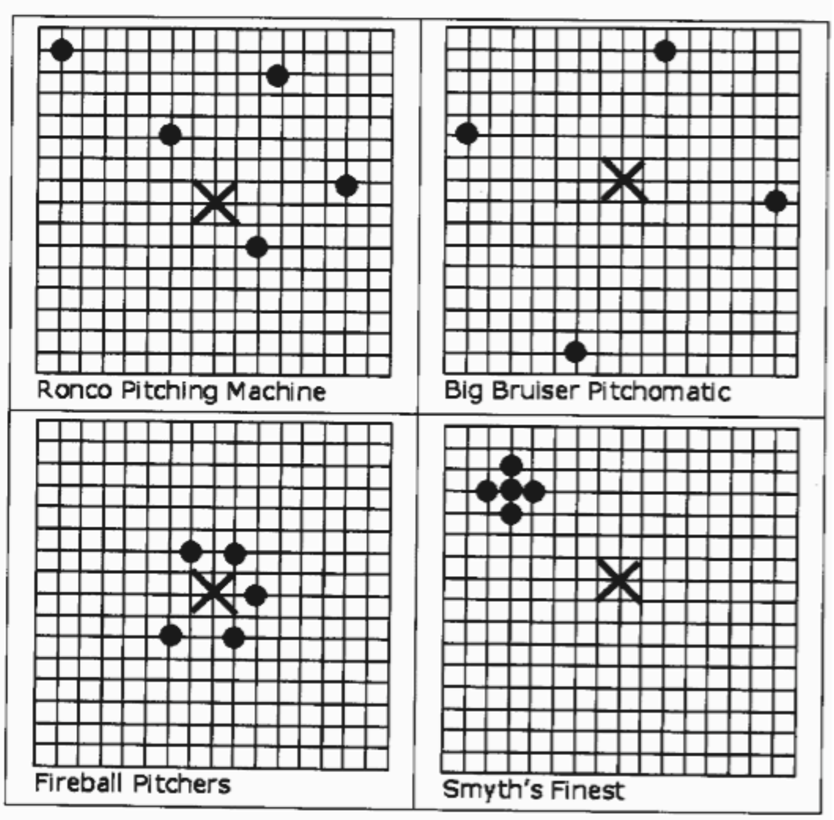
Please explain or give an example

b. Can the MAD ever be negative? Yes / No

Please explain or give an example

Pitching Machines

Here are four grids showing the results form four different pitching machines. The X represents the target and the black dots represent where different pitches landed.



a. Which pitching machine has the lowest variance?

b. Which pitching machine has the highest variance?

c. Which pitching machine has the best average?

(c) Study 2 (variability)

(i) New strategy items

(1) Proportional MAD - learning resource

Is a MAD of $8 considered high or low? Here is a way to check.
Ruth is a waitress. In the last two evenings she earned $2 and $18.
Her average is $10, and her MAD is $8. As you see, the difference between $2 and $18 is very big. Relative to her average, her MAD is very high.

Shanese is the owner of the restaurant. In the same two evenings she earned $102 and $118. Her average is $110, and her MAD is also $8. For her, the difference between $102 and $118 is not very big. Even though she has the same MAD as Ruth ($8), she does not care about it as much.

Conclusion – to evaluate how important the MAD is, we need to compare it to the average. To do that, we divide the MAD by the Average. This is called Proportional MAD.

Proportional MAD = MAD ÷ Average

The Proportional MAD of Ruth:	MAD ÷ Average = 8÷10 = 0.80
The Proportional MAD of Shanese:	MAD ÷ Average = 8÷1,010 = 0.07

The Proportional MAD of Ruth is much higher. This shows that the MAD is more significant compared to her average.

Dennis, the cook, has average daily earnings of $20 and MAD of $4.

What is the Proportional MAD of Dennis?

(2) Proportional MAD - new-strategy item

Watching television

Michael and Rasheed measured how long they watch television every night.
Michael watches on average 40 minutes per night and his MAD is 10 minutes
Rasheed watches on average 100 minutes per night and his MAD is 20 minutes
Considering their averages, estimate which MAD is more significant.

1. What is the proportional MAD of Michael?

2. What is the proportional MAD of Rasheed?

3. Which proportional MAD is higher?
   a. Michael
   b. Rasheed
   c. They are the same
   d. Cannot tell from these numbers

(3) Ignore small errors - learning resource

Sometimes it is okay if machines are not 100% accurate.
For example, if we buy a 5 Lb bag of potatoes, it is okay if the weight is slightly off (since potatoes are never exactly 5 Lb). We still do not want the weight to be very wrong.

Here is how we calculate the MAD if we think that small errors are okay:

Step 1: Find the average.
Step 2: Find the distances from the average.
Then we do something new: If the distance is 1 or less, make it 0, since we do not care about small mistakes.
Step 3: Find the average of the updated distances

Here is an example. If three bags of potatoes weight 3, 6 and 6 Lb, this is how we calculate the new measure:

Step 1: Find the average:	(3+6+6)÷3 = 15÷3 = 5
Step 2: - Find the distances from the average:	From 3 to 5 = 2 From 6 to 5 = 1

- If the distance is 1 or less, make it 0:	From 6 to 5 = 1 2, 1, 1 2, 0, 0
Step 3: Find the average distance	(2+0+0)÷3 = 2÷3 = 0.6

Question 1: Apply the same method. What is the updated MAD of bags of potatoes that weight 2, 6, and 7 Lb, if I do not care about mistakes of 1 Lb or less?

(4) Ignore small errors - new-strategy item

Sony's New Laser

Sony is testing a new laser that can measure distances. It will be a safety feature that measures distances between driving cars. The measure does not have to be accurate. In fact, they are willing to accept mistakes of 1 foot or less.

They tested two lasers at the same distance. Each laser was tested 4 times.
- Laser 1 gave: 2 feet, 7 feet, 7 feet, and 8 feet.
- Laser 2 gave: 3 feet, 5 feet, 6 feet, and 10 feet.

1. If mistakes of 1 foot or less are okay, what is the updated MAD of Laser 1?

2. If mistakes of 1 foot or less are okay, what is the updated MAD of laser 2?

3. Which laser is more accurate?
   a. Laser 1
   b. Laser 2
   c. They are the same
   d. There is not enough information to answer this.

(ii) Other items

Calculate MAD

What is the MAD of the following numbers?

1. (2, 4, 7, 3)

2. (13, 14, 18)

New Pizzeria

Dave and Jamie went to a new pizzeria. They tried 4 different pizzas, and gave them scores from 1 to 10 (10 being the best).
The average score Dave gave was 7 and his MAD was 2
The average score Jamie gave was 4 and her MAD was 2 as well.

Which of the following conclusions is correct?

a. Overall, Dave enjoyed the pizzas more than Jamie did.
b. The pizzas that Dave liked were similar to the pizzas that Jamie liked.
c. Both a. and b. are correct.
d. Both a. and b. are wrong.

Are these methods correct?

1. Last year, Jacob came up with the following method for calculating MAD for these sets of numbers:	A: (3, 3, 4, 5, 10)	B: (1, 6, 8)
Step 1: Find the average	5	5
Step 2: Find the distances from the average	2 2 1 0 5	4 1 3
Step 3: Add up all the distances	2+2+1+0+5 = 10	4+1+3 = 8

What do you think about this method?

a. It works.
b. It does not work because there is different number of numbers in each set.
c. It does not work because the method does not use all the numbers.

2. Marlene came up with a different method for calculating MAD for the same sets:	A: (3, 3, 4, 5, 10)	B: (1, 6, 8)
Step 1: Find the average	5	5
Step 2: Find the distance between the highest number and the average	From 10 to 5 = 5	From 8 to 5 = 3
Step 3: Divide by how many numbers there are	5 ÷ 5 = 1	3 ÷ 3 = 1

What do you think about this method?

a. It works.
b. It does not work because there is different number of numbers in each set.
c. It does not work because the method does not use all the numbers.

Which bus to take?

[a1]

I have a meeting on Friday at CMU at 12:30. Here is the data about the buses I can take:
- 68G: Average arrival time at CMU: 12:21. MAD: 4 min
- EBO: Average arrival time at CMU: 12:15. MAD: 18 min
- 61B: Average arrival time at CMU: 12:29. MAD: 2 min

1. Which bus is most likely to bring me on time for my 12:30 meeting? (circle one)

68G / EBO / 61B / not enough information

2. Someone just told me that the meeting is actually at 12:10 (and not 12:30). Which bus is most likely to arrive at CMU before 12:10?

68G / EBO / 61B / not enough information

Who is more likely to score?

Tim and Brady play basketball on the school team.
Tim’s average points per game is 7.5, and his MAD is 6.
Brady’s average points per game is also 7.5, and his MAD is 2.

Which player is more likely to score more than 7.5 points in the next game?

a. Tim
b. Brady
c. They are equally likely to score more than 7.5 points
d. There is not enough information to answer that

Variability

What does variability measure?

a. How spread the numbers are
b. How big the average is
c. How high the middle number is
d. How large the range is

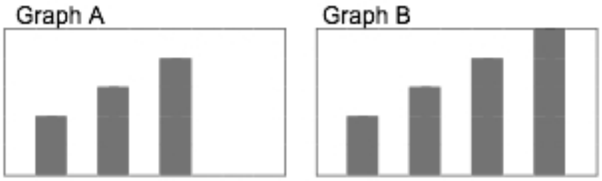
Montreal <-> Vancouver

The average temperatures in Montreal and Vancouver, two cities in Canada, are similar, but the MAD is much higher in Montreal. What do you think it means? (circle one)

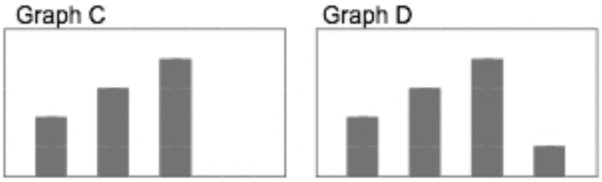
a. Montreal is always warmer
b. Vancouver is always warmer
c. The temperatures in Montreal are more extreme
d. The temperatures in Vancouver are more extreme

Which MAD is lower?

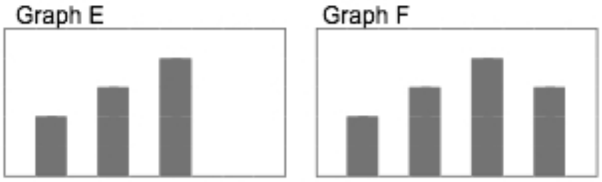
For each of the pairs below, answer which MAD is lower (in which graph the are the bars closer together?).



a. Graph A has lower MAD
b. Graph B has lower MAD
c. They have the same MAD
d. Can't tell without numbers



a. Graph C has lower MAD
b. Graph D has lower MAD
c. They have the same MAD
d. Can't tell without numbers



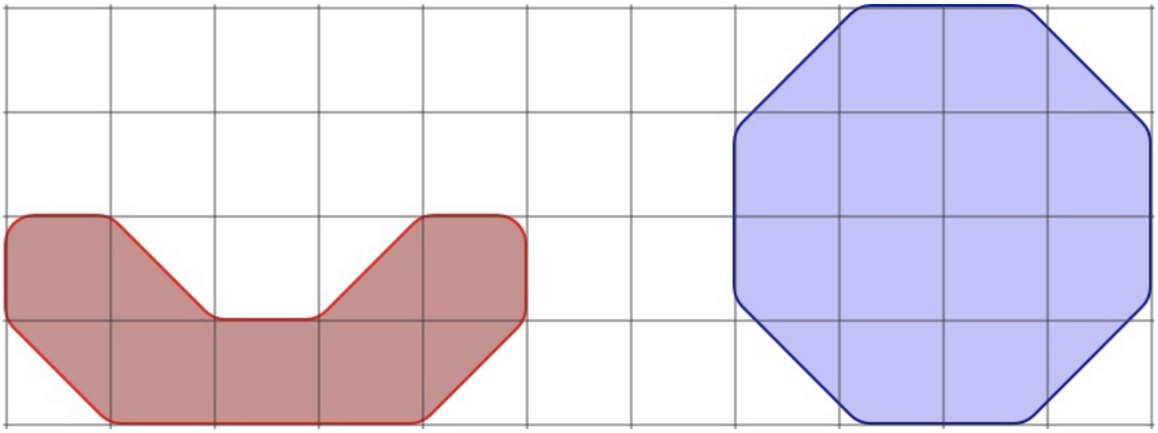
a. Graph E has lower MAD
b. Graph F has lower MAD
c. They have the same MAD
d. Can't tell without numbers

Help Danny

Danny wants to buy a swimming pool for his back yard. He wants to buy the biggest pool he can afford (that is, the pool that can hold the most water). The problem is that the shapes of the swimming pools are weird.

He is considering one of the following two pools:

The Great Bean		**Bee Hive**	
Perimeter:	12 feet	Perimeter:	12 feet
Depth:	2 feet	Depth:	1.5 feet



Danny used the following method to estimate how much water the pool can hold: Perimeter * Depth

The Great Bean:	Bee Hive
12 * 2 = 24	12 * 1.5 = 18

Danny concluded that The Great Bean is bigger than the Bee Hive.

1. Do you think that Danny's method is correct?
   a. Yes.
   b. No.
   c. There is not enough information to know that.

2. Can you think of a better method? Please detail it below:

Method for The Great Bean	Method for Bee Hive
Size of The Great Bean:	Size of Bee Hive:

9.4 The adapted MSLQ

students receive the following questionnaire twice during the studies: before the pre-test and right before the last post-test. Students were asked to report the correctness of each statement for them on a 7-point Likert scale.

The original items were taken from the MSLQ as described by Pintrich (1990, 1993) I have reduced the number of items and included only the following categories in the final questionnaire: self-efficacy, intrinsic value, test anxiety, and self-regulation (cognitive strategies was the only category to be left out). One item was added to the self-regulation category. Another item was added, and together with existing two questions created the new category of liking of math. All categories but test anxiety included three items.

Items are given here according to their category; they were mixed in the test form. Items marked by * are not part of the original MSLQ. Items marked by (r) are reversed, that is, higher response means more negative reaction. Items marked by # were taken from Mitchell 1992 and focus on aspects of situational interest.

(a) Surveys in study 1

Self-Efficacy

Usually I can understand the ideas taught in math class.
I can do a good job in math class.
I think I will receive a good grade in math class this class.

Intrinsic Value

I like to have challenges and to learn new things in Math.
Math is very important to me.
I think I will use what I learn in math later in life.

Test Anxiety (pre-test only)

I am so nervous during a test that I cannot remember anything.
I worry a lot about tests.

Math liking

I like what I am learning in math.
We learn interesting things in math.
* Math is one of the most boring subjects in school (R)

Self-Regulation

When work is hard I either give up or study only the easy parts. (R)
When the teacher is talking I often think of other things and don't really listen. (R)
* I check that my answers make sense before I say I am done

Perceived benefit

* I enjoyed the last several days more than I usually enjoy math
* The last several days in math class have been totally confusing. (R)
* In the last several days I have learned more than I usually learn in math
* I had to think a lot more during math classes this week

(b) Surveys in study 2

Self-Efficacy

Usually I can understand the ideas taught in math class.
I can do a good job in math class.
I think I will receive a good grade in math class this year

***Test Anxiety* (pretest only)**

I am so nervous during a test that I cannot remember anything.
I worry a lot about tests.

personal interest:

# Mathematics is enjoyable to me.
# I have always thought that studying mathematics in school is boring (*R)
# Compared to other subjects, mathematics is exciting to me

Situational Interest (posttest only)

# This study was more fun than a regular class.
# During the study I enjoyed doing math more than usual.
# We did not do anything interesting in math during the study.

Effort (posttest only)

* Math periods during the study required me to do more.
* I had to think a lot more in math during the study
* Math classes during the study required less effort than they usually do (R)