The Invention Support Environment: Using metacognitive scaffolding and interactive learning environments to improve learning from invention

by

Natasha Grace Holmes

B.Sc. (Hons), The University of Guelph, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Physics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

October 2011

Abstract

Invention activities are discovery-learning activities that ask students to invent a solution to a problem before being taught the expert solution. The combination of invention activities and tell-and-practice methods has been shown to lead to better student learning and performance on transfer tasks, as compared to tell-and-practice methods alone. A computer-based interactive learning environment, called the Invention Support Environment (ISE), was built using Cognitive Tutor Authoring Tools to improve the in-class use of invention activities, and act as a research tool for studying the effects of the activities. The system was designed to support three levels of metacognitive scaffolding, using domain-general prompts. It also features a platform for creating new invention tasks within the system, requiring little to no programming experience. The ISE was used to evaluate how domain-general scaffolding of invention activities can best support acquisition of domain knowledge and scientific reasoning skills. Five invention activities in statistics and data-analysis domains were given to 134 undergraduate students in a physics lab course at the University of British Columbia. Students either received guidance in the form of faded metacognitive scaffolding or unguided inventions. It was found that faded metacognitive scaffolding did not improve learning of invention skills compared to unguided inventions. Faded metacognitive scaffolding was found to improve understanding of domain equations, as seen through higher performance on debugging items in a statistics diagnostic. Future experimental design and ISE improvements are discussed.

Preface

While this research uses human subjects during an in-class intervention, the purpose of the research was to quantitatively evaluate in-class activities for the purpose of institutional research. The research results do not identify any individuals involved in the study.

Table of Contents

Appendices

List of Tables

List of Figures

List of Abbreviations

ANCOVA Analysis of Covariance

BRD Behavior Graph

CDPA Concise Data Processing Assessment

CTAT Cognitive Tutoring Authoring Tools

FG Faded Guidance

ILE Interactive Learning Environments

IPL Invention as Preparation for future Learning

ISE Invention Support Environment

ITs Intelligent Tutoring Systems

PER Physics Education Research

PF Productive Failure

TA Teaching Assistant

UBC University of British Columbia

UI Unguided Invention

Acknowledgements

I would like to offer sincere thanks and gratitude to a number of individuals who have assisted me through the past two years.

To Dr. D. Bonn, who has given me the opportunity to study a field that I am truly passionate about. To Dr. I. Roll for getting me acquainted with the world of education research, pushing me to take control of my project, and providing me with professional opportunities and insights. To Dr. J. Day for playing the devil's advocate, and for keeping me grounded in physics.

To Anthony Park for his assistance with data analysis over the summers, and for putting up with me as I learn how to supervise.

Special thanks goes out to Tim and Ricardo at Explora Comm. Inc. for their wonderful work creating the Equation Editor and Upload components. This was an integral part of getting inventions on the computer. I am infinitely grateful for their efficiency, creativity and diligence.

And of course this study would not have been possible without the technical support from the folks at CTAT (especially John and Octav) and the UBC IT Staff (especially Ron Parachoniak).

This work was supported (in part) by the Pittsburgh Science of Learning Center which is funded by the National Science Foundation, award number (#SBE-0836012), and by the University of British Columbia through the Carl Wieman Science Education Initiative.

Chapter 1

Introduction

Traditionally, undergraduate physics courses involve a variety of learning opportunities such as lectures, homework assignments, tutorials and lab activities. Lab activities are often seen as a chance for students to apply the theories they have learned in lecture to real-world situations and scientific experiments. Several studies in Physics Education Research (PER), however, have suggested that students struggle to connect real-world physics with that in the classroom. For example, students believe that gaining a deep understanding of physics and getting good marks in a physics course are not the same and require very different actions in order to be achieved [20]. Use of the Colorado Learning Attitudes about Science Survey has demonstrated that while many students are aware of physicists beliefs about physics, they do not find these beliefs to be relevant to their own learning [21]. This distinction between physics knowledge acquired in the classroom and physics in practice suggests that traditional physics courses are not connecting lessons in the classroom to real-world physics inquiry.

In traditional undergraduate physics lab experiments, students are typically given a series of detailed instructions in order to conduct an experiment to observe some physical phenomenon that has been previously studied. The 'cookbook' style support results in little to no thinking done by the student, and does not require them to reason through the experiment for themselves. Instead these labs ought to help prepare students to become competent physicists who can investigate new questions with creative, innovative approaches, rather than to simply follow instructions. Inquiry-based (or discovery-based) learning activities are structured tasks that support students through metacognitive, reasoning and innovative inquiry behaviour rather than procedural, domain-specific actions. These types of tasks have been shown to support students in acquiring reasoning skills, while also supporting domain learning [5] [19] [44]. For this thesis, I examine the use of invention activities, a type of inquiry-based learning activity, in an undergraduate physics lab.

1.1 Invention Activities

Invention activities are discovery-learning activities where students are asked to invent solutions to complex problems before they have been taught the expert solution. These inventions can be of various forms, including mathematical models, graphing techniques or complex physical systems. In this thesis, a standard invention activity is one in which the goal is to invent a mathematical formula that can be used across cases to solve the task at hand [35] [39]. These tasks set a clear goal and are supported by contrasting cases upon which students base their inventions, with contrasts between pairs that demonstrate the important features of the domain. Execution of the activities typically involves student collaboration and peer instruction [2] [18]. These features will be explained in more detail in the following sections.

Invention activities, when used in a specific framework, have been shown to lead to better learning and performance on near- and far-transfer problems than through traditional direct instruction [8] [33] [35] [39] [42]. This framework, known as invention as preparation for future learning (IPL) [39], involves 3 steps. First, students work through an invention activity in pairs or small groups. Students are then provided with direct instruction on the expert solution to the task. Finally, they are given an opportunity to use the expert solution on several practice problems. An interesting result of these tasks is that students demonstrate higher level performance on transfer tasks even though they typically fail to invent the correct solution. This phenomenon, which has also been observed in the context of ill-structured physics and math problems, has been referred to as Productive Failure (PF) [24] [25] [26] . In these studies, students who worked through challenging problems with little support produced poorer quality solutions than those students who received support. These unsupported students, however, outperformed the supported ones when it came to near- and far-transfer problems.

Several studies have criticized discovery-learning activities demonstrating situations where direct instruction methods are superior to discovery learning [27] [40] [41]. The discovery-learning activities used in these studies often involve knowledge being withheld from the students. The IPL and PF frameworks, in contrast, depend on a discovery-learning period that is necessarily followed by direct instruction, or a consolidation phase, which explicitly provides the relevant knowledge [2]. With an invention task prefacing any instruction on the topic to be introduced, the IPL framework demonstrates that inventions prepare students to learn from direct instruction, but do not replace the instruction. In this case, students prepare themselves to receive the content knowledge by engaging in high-level reasoning behaviour and examining the full solution space of the problem. It is this engagement in scientific reasoning that allows students to solve higher-order transfer or application items, outside of the activities explicitly taught [25] [39].

It has been shown that when constructivist instruction produces lower learning gains than direct instruction, students are attempting to learn domain content while also engaging in new reasoning behaviours with insufficient support [41]. That is, they reach a cognitive overload as they attempt to learn new thinking strategies in order to learn a new and challenging concept. Structuring the discovery environment, however, may help support the use of novel reasoning skills and reduce this cognitive load. Assisting students with the reasoning behaviours would leave more opportunity for students to study the task at hand. A review of discovery-learning with simulation activities, however, demonstrated that it was not yet clear how or when to use scaffolding in discovery-learning activities [19]. The next section describes research done to examine various types of scaffolding for discovery-learning activities and invention activities in particular.

1.2 Structure of Invention Activities

Structuring any type of problem solving activity can be done by providing domain-specific or domain-general support. Domain-specific support refers to support for learning the content material whereas domain-general support refers to supporting the use of reasoning behaviours, which will, presumably, allow the student to discover the content. While domain-specific support in science problem solving tasks has been shown to lead to better content knowledge, domain-general

scaffolding supports development of scientific reasoning skills such as planning, monitoring and evaluation [10]. It has also been shown that domain-independent metacognitive scaffolding in invention activities leads to better student inventions and higher quality reasoning that focuses on the key features of the problem [36]. Invention activities include minimal, implicit support or explicit metacognitive scaffolding.

1.2.1 Explicit Scaffolding

A series of metacognitive prompts, such as domain-independent questions and self-explanation, demonstrates the use of explicit scaffolding. Self-explanation, where learners form written statements to explain their reasoning and strategies, have been shown to enhance learning gains, performance on transfer-tasks and integration of new information [4] [14]. In invention tasks, students have been found to spontaneously self-explain their invented solutions, and those who receive scaffolding during invention tasks produce higher quality self-explanations during the design phase [36].

The particular behaviour prompts, while specific to the stage of the task, are kept independent of the domain and can be recycled for use in other tasks of similar structure. The scaffolding was developed to engage students in reasoning behaviours that were considered useful by experts for inventions. The framework for the scaffolding involves five stages: task definition; analysis; planning and design; implementation and interpretation; and evaluation [36].

1. *Task Definition*

   Students first need to understand what they are being asked to do through clear descriptions of the context and goals of the task. To achieve this, students are provided with a short introduction, and data on which to base their inventions.

2. *Analysis*

   Students analyze the data by making comparisons between pairs of graphs and explaining qualitative differences, thus extracting features from the contrasting cases. In order to ensure students examine all possible comparisons, students can rank all the data sets according to the index they are being asked to invent. In contrast, a single, general explanation across all cases has also been found to be useful for extracting underlying concepts that generalize across the multiple cases [12].

3. *Planning and Design*

   At this stage, students begin to develop the mathematical representation to quantitatively assess the features of the domain. Ideally, this process follows straightforwardly from the analysis phase, where the features are directly linked to an algebraic definition.

4. *Implementation and Interpretation*

   Using the data provided and the designed formula, students implement their inventions for each of the data sets. The results are interpreted through a final quantitative ranking.

5. *Evaluation*

   With the initial qualitative and final quantitative rankings, students evaluate the success of their inventions. In addition, self-explanation of their methods after implementation allow them to reflect on conceptual or computational challenges to their formula.

1.2.2 Implicit Scaffolding

While an initial attempt at an invention ought to begin with task definition and analysis, the first level of implicit scaffolding is that all subsequent stages are iterative and cyclic. This feature should be implicitly supported and encouraged throughout invention, so that students understand the cyclic nature of scientific reasoning, for example by making it easy to freely move back and forth between sections of the task.

Peer interaction is another example of implicit scaffolding, for example by having students complete invention tasks in small groups or pairs. The use of peer interaction during invention activities elicits feedback between students and allows them to obtain assistance in identifying gaps in reasoning. While students rarely interact with other groups [36], valuable discussion can still occur within groups or between pairs.

Another necessary example of implicit scaffolding is the way in which the invention activity data is presented. In order to highlight features of the domain, the data can be presented as a series of contrasting cases, in which pairs of data can be found that differ by only a single domain feature. This allows students to notice what about the data is relevant to solving the task [38]. In addition, students use the contrasts to discern what effect the varied feature has on the final result.

The "Fuel Consumption" activity, for example, which was used by Roll, et al. (accepted) [36], describes four car companies that tested the fuel efficiency of a single vehicle through measurements of distance traveled with various amounts of fuel. The activity was created to precede a lesson on determining the uncertainty in the slope of linear data going through the origin. Contrasts between Contractor A and subsequent cases in Figure 1.1 allow the user to extract the three important features through the following comparisons:

- A vs B: Increasing the position of the measurements further along the x-axis will decrease the uncertainty in the slope, a feature hereby referred to as "leverage."
- A vs C: Increasing the number of measurements taken will decrease the slope uncertainty, a feature hereby referred to as "sample size."
- A vs D: Increasing the spread between the data and the line will increase the slope uncertainty, a feature hereby referred to as "residual distances."

From these contrasts, a student might discern that a formula for the uncertainty in the slope should be inversely proportional to the average horizontal distance, inversely proportional to the number of measurements taken and proportional to the sum of the squared residual distances between the points and the line. This reasoning could get one close to the actual mathematical solution to this problem, represented in Equation 1.1:

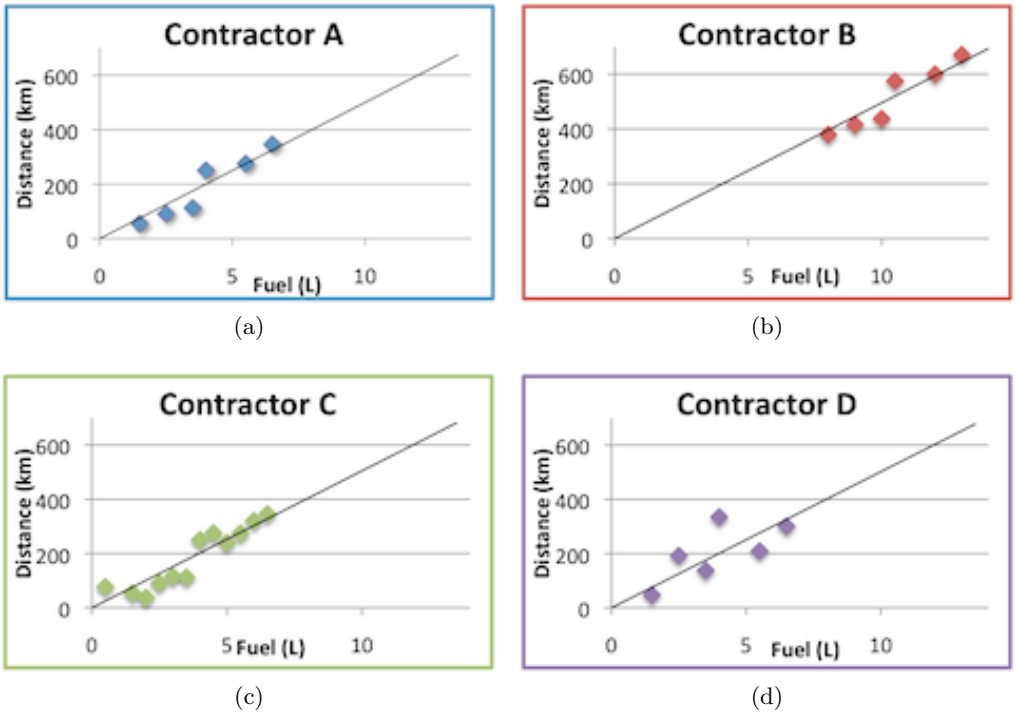


Figure 1.1: The data for the fuel consumption invention activity shows a set of contrasting cases. Each graph shows the distance traveled by various cars with various amounts of fuel, in order to determine the fuel efficiency (the slope of the graph). The uncertainty in the fuel efficiency is, therefore, the uncertainty in the slope of the best fitting line with a zero intercept.

$$\delta m^2 = \frac{1}{N} \frac{\frac{1}{N}\sum_{i=1}^{N}\Big(y_i - f(x_i)\Big)^2}{\frac{1}{N}\sum_{i=1}^{N} x_i^2} = \frac{1}{N} \frac{\sum_{i=1}^{N}\Big(y_i - f(x_i)\Big)^2}{\sum_{i=1}^{N} x_i^2}, \tag{1.1}$$

where m is the slope of the line, δm is the uncertainty in the slope, N is the number of points in the data set, y_i is the y-value of the i^{th} data point, x_i is the x-value of the i^{th} data point, f(x_i) is the y-value on the line that corresponds to x_i (that is, f(x_i) = m x_i).

1.2.3 Faded Scaffolding

This thesis attempts to further develop an understanding of how invention activities support student learning and reasoning by examining the effects of different ways to structure the tasks. While student engagement with the domain-general scaffolding improves performance in the task itself[36], it has not been demonstrated whether the scaffolding helps students acquire the particular skills presented. That is, the behaviours that were used proved to be useful for inventing better solutions, but it is unclear whether the method of supporting the strategies allows students to learn to use them on their own. In addition, neither measures of learning or transfer have been examined with this type of scaffolding.

In order to observe whether or not students could use these skills without prompting, measure-

ments could be made of behaviour before and after removal of levels of domain-general support. Fading levels of domain-general support have been shown to produce equal performance on problem solving tasks as continuous domain-general support [10] This suggests that students fading domain-general support allows students to internalize the metacognitive strategies so that they can be used without prompting. Fading support over time has also been recommended for students to develop their own self-regulating strategies and become independent learners [7].

Therefore, the primary research questions for this thesis are:

1. How can a series of invention activities be structured to improve student development of scientific reasoning skills while also improving learning at the domain-level?
   (a) Does faded-scaffolding of invention activities improve student development of *scientific reasoning skills* when compared to unguided invention activities?
   (b) Does faded-scaffolding of invention activities improve *domain-level knowledge* when compared to unguided invention activities?

A significant challenge to overcome when studying these questions in a classroom (compared to in a controlled lab setting) is to ensure the equivalence of the experiment conditions, especially with regards to delivery of invention tasks. Use of a computer system or interactive learning environment (ILE) could help overcome this challenge as well as provide additional benefits of being scalable, generalizable and transferable between tasks, classrooms and instructors.

1.3 Interactive Learning Environments

Currently, the use of invention activities in classrooms is not very widespread, and are mostly used by experts who study their effectiveness. Instructors may find it challenging to witness their students struggling with the activity, especially if they are unfamiliar with the research-based outcomes of the tasks. Informal classroom observations showed that instructors may provide too much support while students develop inventions. This extra support tended to block students' thinking during the invention process. The approach I have taken to address this issue is to make the activities entirely self-contained through the use of interactive learning environments (ILE).

An ILE is a computer-based system that supports students in learning a concept through interactive engagement with the computer interface. Often, ILE are used for repetitive practice or highly-supported learning environments [28] [43], compared to discovery-based learning. However, Azevedo and Hadwin (2005) [7] describe five studies that examine how computer-based learning environments can be used to scaffold for self-regulated learning and metacognitive strategies, rather than procedural knowledge [6] [15] [16] [22] [32]. Nonetheless, few attempts have been made to apply this technology to inquiry-based problems such as invention activities.

One such attempt was through "The Invention Lab [34], an intelligent tutoring system (ITS) for delivering an invention task about statistical variability to middle school students. An ITS is a type of ILE that specifically tries to reproduce learning gains seen through human tutoring. They usually involve adaptive feedback that is tailored for particular student actions, patterns of behaviour, or progress in learning. In fact, the Invention Lab adapted tasks to address particular

misconceptions identified in previous problems. In particular, the system described two trampoline companies and a series of measurements of how high balls bounced on their trampolines. Students were asked to determine which company they thought was more consistent in their bounces (less spread out) and then to invent and implement a method to quantitatively evaluate each trampoline. Finally, comparisons of their initial predictions and the results of their implementations allowed students to evaluate their invented methods.

While students who used the system created models and developed inquiry skills, the stepwise method of creating models seems to promote exploratory analysis behaviours, similar to those observed in novice problem solvers [37]. That is, when creating a model the user would immediately apply their ideas to the data sets and receive a numerical result without a general model visible (that used variables as in a mathematical equation, for example), supporting a trial-and-error style of development. They also received immediate feedback based on their initial comparison (ensuring students make correct predictions) and if their final ranking did not match the initial, correct prediction. While this may be necessary support for middle-school students, the immediate feedback, and step-by-step inventing does not support the high-level reasoning that is expected from university students. In order to support more expert-like problem solving and reasoning behaviours at the college-level, an invention interface should allow users to create more complex, general mathematical models with minimal feedback.

There are several benefits for use of ILEs with invention activities. In the classroom, an ILE for inventions could make these tasks more accessible to instructors who have little to no prior knowledge of the IPL framework. Example-Tracing Tutors using Cognitive Tutor Authoring Tools [3] [29], for example, allow instructors with no programming experience to manipulate particular tasks using a graphical user interface. Activities can be easily graded through logging features and also assist in efforts to produce "paperless" labs.

In addition, ILEs have several benefits for use in research settings. They make randomized experimental conditions manageable within classrooms, whereas paper and instructor-led tasks are typically used with convenience sampling across classes [36]. Automatic logging capabilities provide time-stamped records of student actions in the system for evaluation and study. These logs produce fine-grained data that allow student reasoning and behaviour patterns to be examined.

With this motivation, we have designed an online ILE called the Invention Support Environment (ISE) to deliver invention activities in a classroom. This introduces further research questions:

1. Can an interactive learning environment be built for invention activities?
   (a) Would an ILE for invention activities improve how students learn through invention?
   (b) Can an ILE be built for invention activities that scaffolds successful metacognitive strategies for inventions?
   (c) Can an ILE be built for invention activities that will assist with dissemination of the pedagogy and be easily adapted to new invention tasks?

This thesis is divided into 7 chapters. Chapter 2 describes the ISE interface and how invention scaffolding is supported. Chapter 3 focuses on details and descriptions of the ISE and how it was

built. Using the ISE, an in-class experiment was conducted to examine the effects of the scaffolding. Chapter 4 details the specifics regarding the experimental design and methods. Chapters 5, 6 and 7 present the results and conclusions of three different assessments (a statistics diagnostic test, a near-transfer invention task about the independent t-test and a task to recreate data from the t-test activity, respectively). Chapter 8 reviews the information presented in the thesis.

Chapter 2

The Invention Support Environment

This chapter presents the user-side of the Invention Support Environment (ISE). Section 2.1 describes how key invention activity structures are supported in the system. Section 2.2 summarizes the system and provides suggestions for future directions.

2.1 Scaffolded Invention Activities

To encourage peer-interactions, students work in pairs at a computer when completing an invention activity. Each pair of students receives a login ID that they use to access each activity. Once logged onto the system, students could select the invention task, submit their student IDs and begin completing the activity.

Key invention strategies were scaffolded using input prompts associated with the behaviours defined in Section 1.2: Task definition, analysis, planning and design, implementation and interpretation, and evaluation. Table 2.1 outlines the activity-general prompts and Figures 2.1 to 2.3 demonstrate their use in the fuel consumption activity (described in Section 1.2). These examples refer to a high-level of scaffolding and some of these elements can be removed to create lower levels, such as when fading. Some research done to examine the effects of different levels of scaffolding can be found in Chapter 4.

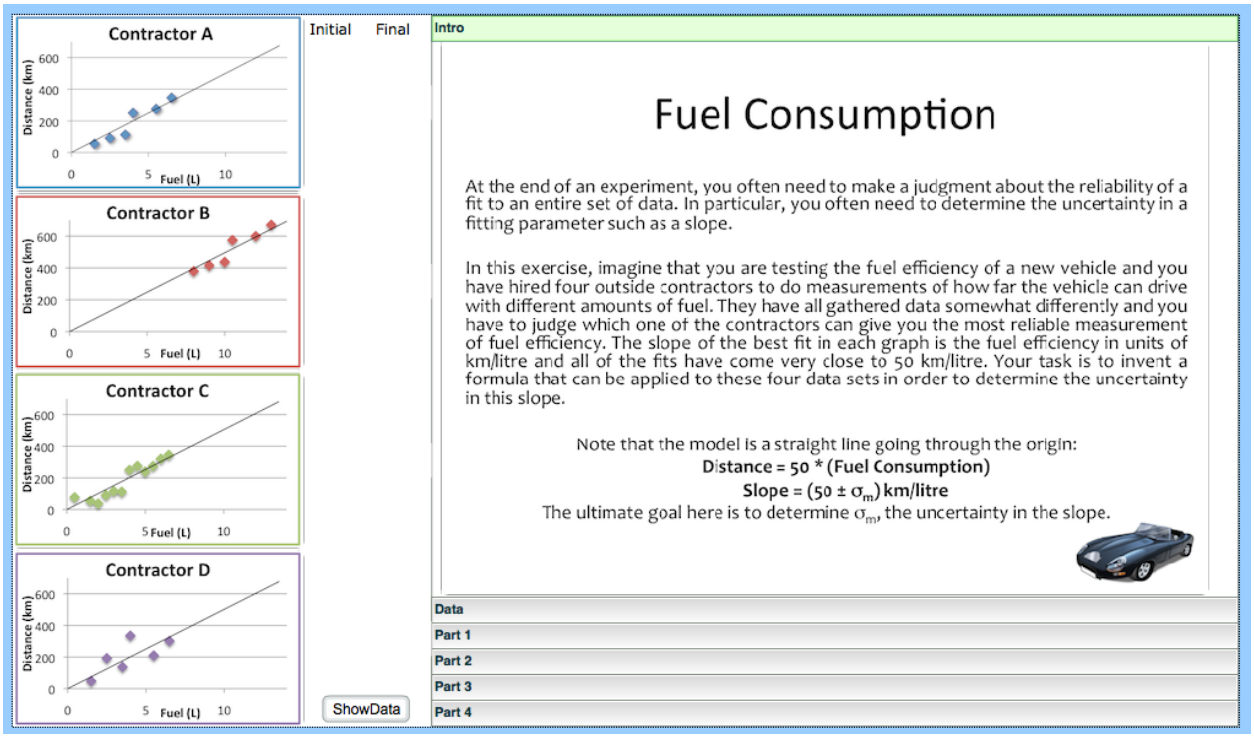
Scaffolding Stage	*Prompt Details*
Task Definition	Introduction Story
Analysis	Explicit pairwise comparisons with self-explanation Qualitative ranking of data sets
Plan and Design	Invent a formula Explain the solution based on task analysis
Implementation & Interpretation	Implement invention for each data set Quantitative ranking of data sets
Evaluation	Compare qualitative & quantitative rankings Explain strengths & weaknesses of invention

Table 2.1: Prompts used for each of the stages of the highest-level of domain-general metacognitive scaffolding during the study.

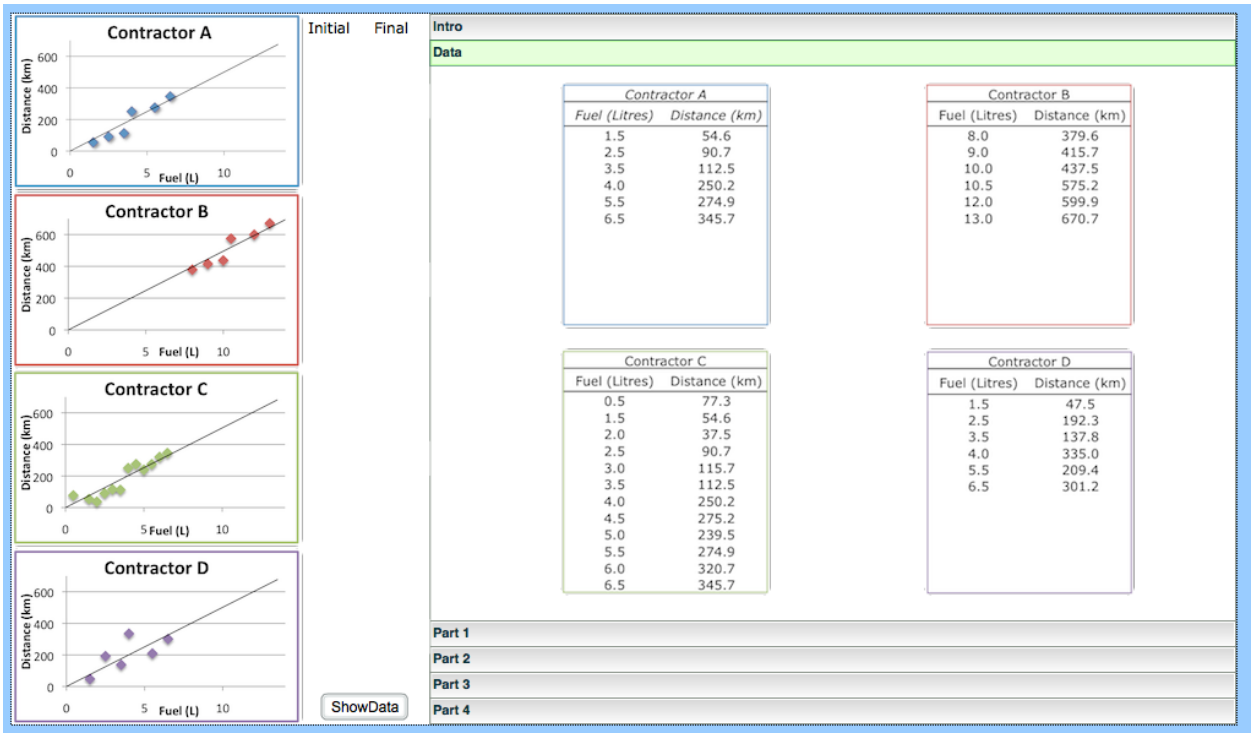
2.1.1 Task Definition

Students receive an introductory story that describes the task and presents the problem they are being asked to solve (shown in Figure 2.1a). Stories are based in non-physics topics and use

cover names for the concepts or variables that students will invent. This minimizes the chance of students entering an 'equation hunting' mode, or simply searching the internet for a solution. It also affirms the notion that there is no correct solution for the activities, since the variables do not exist. Students can then turn to the contrasting data sets, either using the panel on the left side of the main accordion (such that the graphs are always present on the screen) or through an enlarged version accessible through a "Show Data" button towards the bottom of the screen. The next tab in the accordion presents the data tables so users have access to the specific values during implementation (Figure 2.1b). It was decided that visibility of the graphs is much more valuable than that of the tables, since students could extract features and patterns in graphs more easily than in lists of numbers. The tables were, therefore, accessible in their own accordion tab, whereas the graphs would remain visible during the whole activity.



(a) Introduction Story and Graphs



(b) Data and Graphs

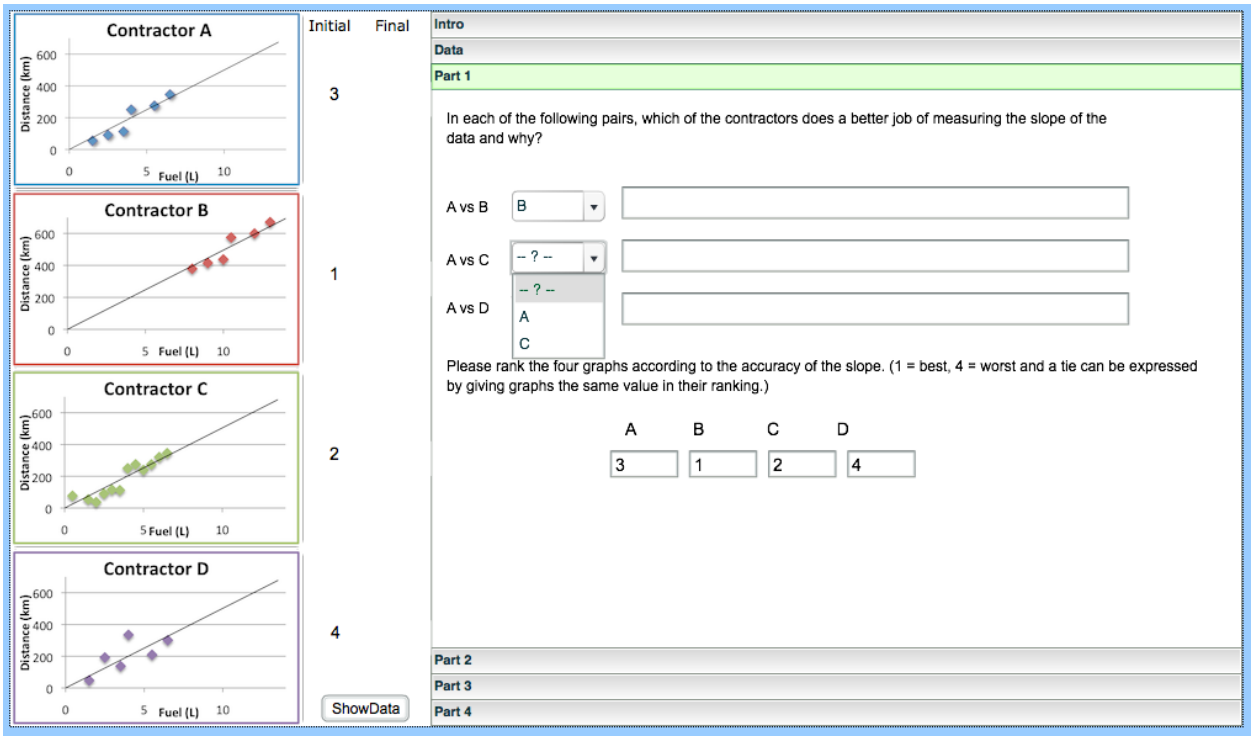
Figure 2.1: Screen shots demonstrating the task definition segments of metacognitive scaffolding for the highest level of support in the fuel consumption task.

2.1.2 Analysis

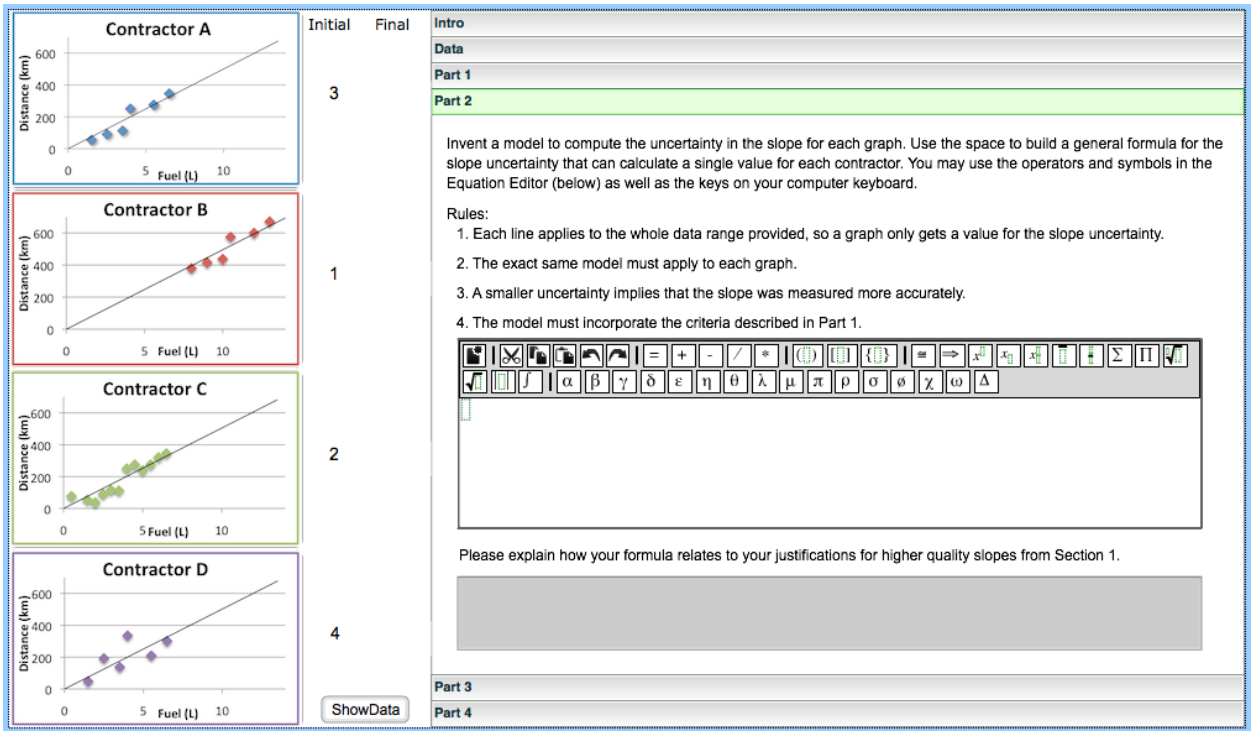
In part 2 of the accordion, Figure 2.2a, students are instructed to compare pairs of data sets and use the drop-down menus to select which, for example, is better at the particular measurement. Prompted self explanations for three pairs of graphs direct students' attention towards the different features across the cases and engage them with the contrasts. In order to ensure students spend time generalizing their comparisons across all provided data, thereby examining the full solution space, students are asked to rank each of the graphs based on the criteria specified for the task. For example, for the fuel consumption activity, they would rank the graphs based on the lowest uncertainty in the slope. This qualitative ranking appeared next to the graphs on the left so that it was visible throughout the rest of the activity. In Figure 2.2a, the user has ranked the graphs BCAD, from lowest to highest uncertainty in the slope. The method of ranking, where students input ranking positions for each dataset, was chosen so that students could assign graphs equal ranking by giving them the same number. This seemed a better alternative to using drop-down menus or assigning datasets to the individual ranking positions, since these methods might unintentionally convince students that datasets could not be equal.

2.1.3 Planning and Design

The next section asks students to invent a single formula that is to be applied across all cases (Figure 2.2b). The equation editor, allows students to build a single, general mathematical formula for their inventions using keyboard input or mathematical symbols and operators in the toolbar along the top. Compared with single-line text, the equation editors acts to visualize the complex mathematical formulas they create on the screen, as though the equation were being written down on paper. The equation editor was built by Explora Communications Inc. and is described in more detail in Section 3.2.1. Next, students then explain how their formula relates to the reasoning made during the previous step through a self-explanation prompt.



(a) Analysis, featuring pairwise comparisons and ranking



(b) Planning and design, featuring the equation editor and prompted self-explanation

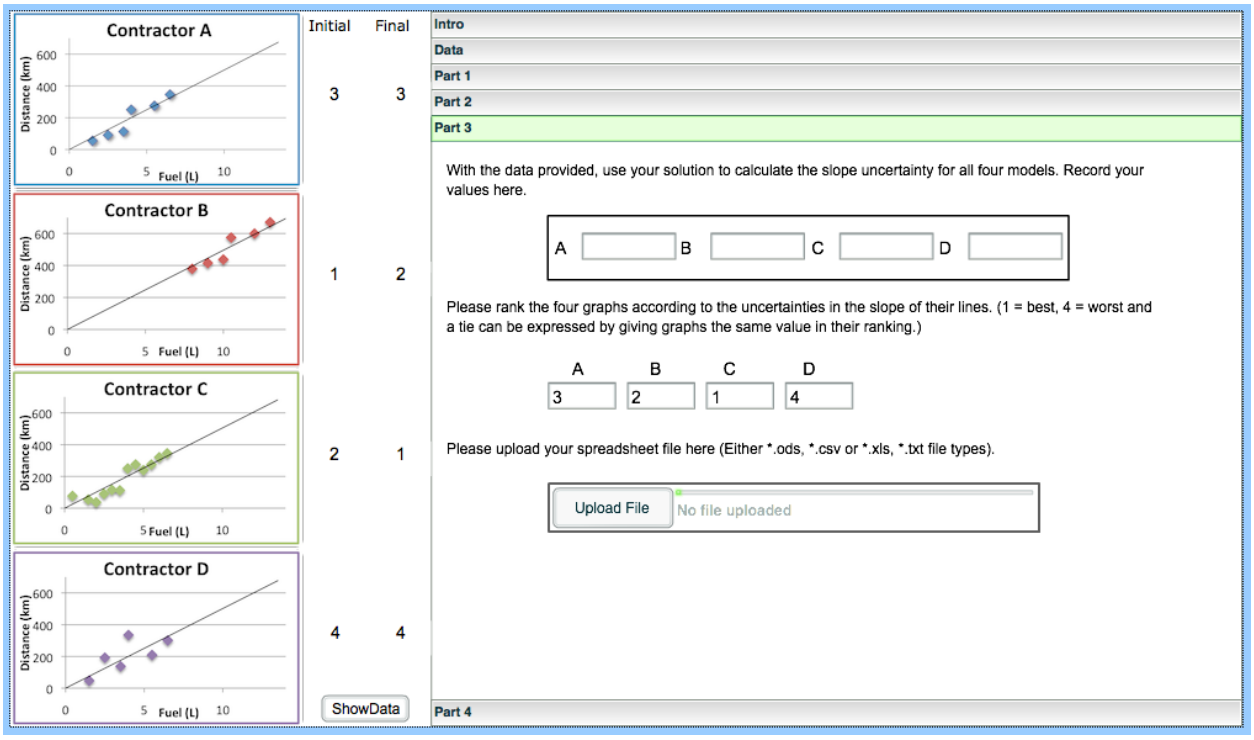
Figure 2.2: Screen shots demonstrating the analysis, and planning and design segments of metacognitive scaffolding for the highest level of support in the fuel consumption task.

2.1.4 Implementation and Interpretation

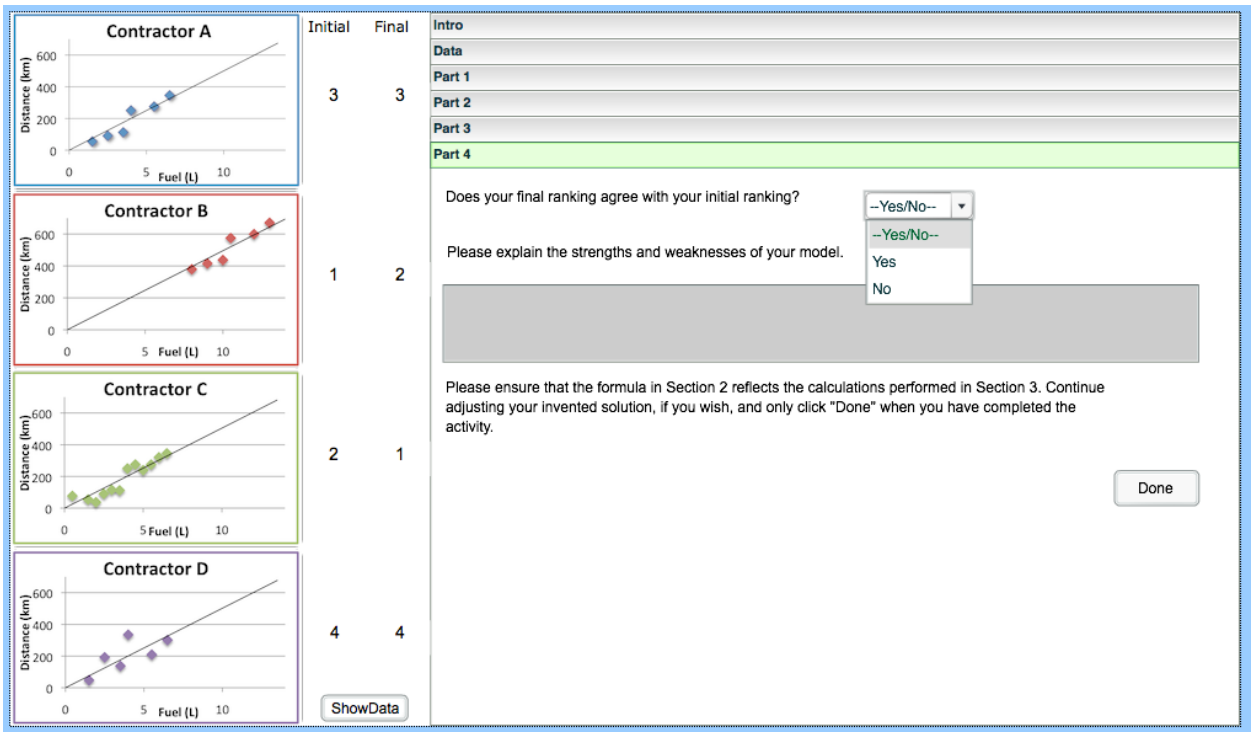
To implement their invented formula using the data provided, students turn to external spreadsheet software. Once calculated, they return to the system to enter their final values in Part 3 of the accordion (Figure 2.3a). An upload feature, also built by Explora Communications Inc., was included so that students could save their spreadsheets to the server (which also houses the activity source files and student log files) to be analyzed separately. A final quantitative ranking is then entered, which would appear next to the initial qualitative ranking, such as CBAD in Figure 2.3a.

2.1.5 Evaluation

With the initial qualitative and final quantitative rankings side by side next to the graphs, students are asked to determine whether the rankings are in agreement. They are then asked to explain why or why not, based on strengths and weaknesses of their invented formulas (Figure 2.3b).



(a) Implementation with upload feature



(b) Evaluation

Figure 2.3: Screen shots demonstrating the implementation, interpretation and evaluation segments of metacognitive scaffolding for the highest level of support in the fuel consumption task.

2.2 Conclusions and Future Directions

This chapter described the Invention Support Environment (ISE), an online interactive learning environment for invention activities. The system supports iterative and cyclic progress through the invention process via an accordion structure, and addresses key invention strategies at each section of the accordion. All of these levels of support are described in detail using the fuel consumption activity as a sample. In addition, generation of complex mathematical formulas using high-level reasoning behaviours, rather than trial-and-error invention, is promoted through the use of prompted self-explanation and an embedded equation editor.

While this chapter describes how invention skills were incorporated into the system, it has not yet been quantitatively compared to paper-tasks. Comparisons should be made to examine student performance when completing inventions on the ISE or on paper. This would determine the effectiveness of the ISE at delivering invention activities in a classroom.

Future versions of the ISE should attempt to make invention activities as self-contained as possible. For example, including a spreadsheet component would ensure students need only access the system itself to be able to complete inventions. Creating computer-based instruction and practice activities would also assist in making the ISE a free-standing software for the full sequence of invent, tell and practice. These items are elaborated further in Section 3.5.

Chapter 3

Technical Features of the ISE

This chapter describes the technical features of the Invention Support Environment (ISE) first presented in the previous chapter. Section 3.1 specifies the development of the ISE interface and system behaviour. Section 3.2 outlines how invention activity scaffolding is supported in the system. Section 3.3 describes the logging capabilities of the ISE, especially for research-purposes. Section 3.4 describes useful features for implementing the ISE in the classroom. Section 3.5 summarizes the technical features of the ISE and proposes various new directions for future development of the system.

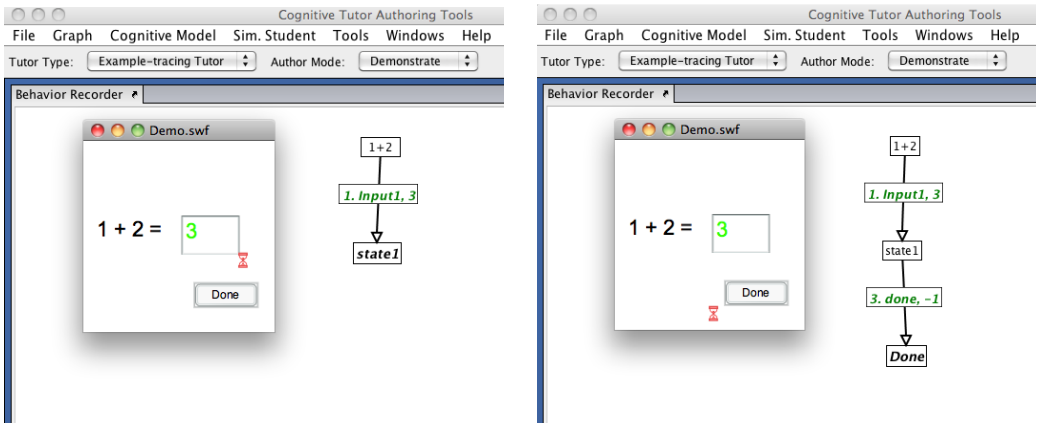
3.1 Building the ISE

To build the ISE, a single interface was programmed in Adobe® Flash® Professional CS5 [1] for all inventions. The interface, that is, the visual structure of the program, was designed to facilitate the invention process through the five scaffolding levels as described in Section 1.2. In general, the interface involved various components, such as images, text areas and buttons, the use of which are described more thoroughly in Section 2.1. The components in the interface were then connected to the behaviour of the tutor using Cognitive Tutor Authoring Tools (CTAT) version 2.10.0 [3] [29].

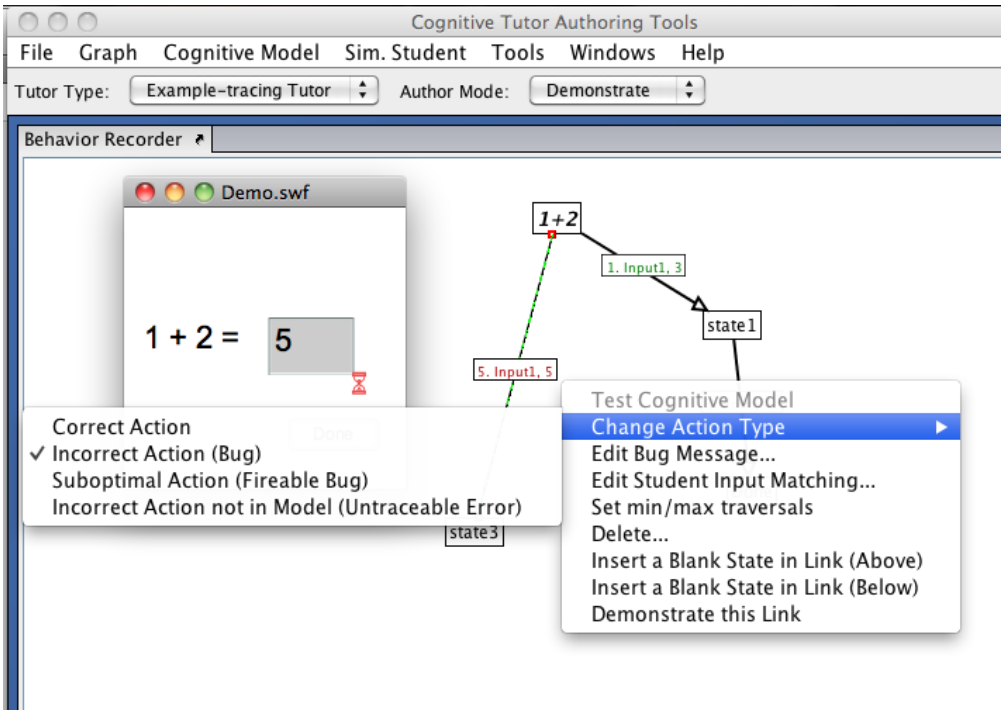
CTAT is a development environment for computer-based learning environments that can provide adaptive feedback to an interface based on cognitive models, interactive behaviours, and loggable actions. CTAT supports two development environments: Cognitive Tutors, programming-intensive ILEs that rely on complex cognitive theory and modelling; and Example-Tracing Tutors, which program by example or demonstration in an interface, requiring no programming background [3]. Example-tracing tutors were used for the ISE so that the system could easily be manipulated by researchers and instructors in the future, regardless of their programming skills, and since invention activities did not require intellectual behaviour.

3.1.1 The Behavior Graph

Behaviour of the system was developed through the behaviour graph (BRD file), a tree-like graph with 'branches' that represent possible student paths, and individual steps or links corresponding to individual student actions. Individual steps in the BRD can either be input manually through a menu or demonstrated in the interface. For example, Figure 3.1a shows that since the developer has input the value of '3' into the blank text area called 'Input1' as the first correct action, CTAT creates step 1 as assigning 'Input1' the value of 3, thus reaching 'State 1'. Subsequently, as seen in Figure 3.1b, a second state is defined through clicking the 'done' button. A sample BRD for the Fuel Consumption activity can be found in Figure 3.2



(a) (b)



(c)

Figure 3.1: A BRD for an example-tracing tutor being built in demonstrate mode. The developer has demonstrated to the system that a) inserting the number 3 into the text area called 'Input1' should be the first action, and b) clicking the done button should be the second. They have also demonstrated in c) that inputing the number 5 in the text area should be flagged as an incorrect action, or bug.

3.1.2 Features of the BRD

The BRD has various features that can be assigned to individual student actions. Firstly, actions can be deemed 'correct' or 'incorrect.' For example, while Figure 3.1a defined inputing '3' as the first step to be a correct action, Figure 3.1c shows that inputing '5' into the space can be classified as an incorrect action (or a bug). In addition, feedback or hint messages can be created to coincide with particular behaviours, such as when students input incorrect values. This can be done through

the menu option "Edit Bug Message..." seen in Figure 3.2.

The Invention Lab [34] is an example of a previously developed ILE for invention activities for middle-school students, described more thoroughly in Section 1.3. This system provided feedback to students for various actions, such as incorrectly ranking contrasting cases, failing to implement the same method for both data sets or failing to notice if final and initial comparisons did not match. In particular, they relied on the intelligent novice model [31], whereby feedback is delayed only until the student fails to demonstrate error detection or correction skills. For the ISE, however, it was deemed important to allow the students, now at the college level, to use their own monitoring skills and peer interactions for these sorts of actions, rather than obtaining it automatically from the system. That is, feedback should be delayed until students have completed their inventions and had the opportunity to evaluate their designs. Necessary feedback and error correction would instead be identified and discussed during the consolidation phase. No feedback on student actions was therefore provided for any stages of the ISE.

The user's progress along a path can be 'ordered' or 'unordered.' In ordered mode, the student must progress sequentially through each step in the BRD starting at the initial start state. In contrast, the unordered mode allows the user to start with any step and progress between each of the steps in any order. In addition, the number of allowed times a student may traverse a step can be varied, so the student could only attempt an action once, or multiple times, depending on the value set through the "Set min/max traversals" menu option in Figure 3.2.

Components and steps in the BRD with common features can also be grouped into sections. This feature permits behaviour within and between groups to be varied. For example, while progress between groups can occur in any order, the progress through one particular group could be ordered so that users must progress step-wise between the BRD steps. For a system to support invention activities, it must encourage evaluation by allowing students to freely move back and forth between sections of the task. For this reason, all steps in the ISE were unordered and had no upper limit on the number of traversals.

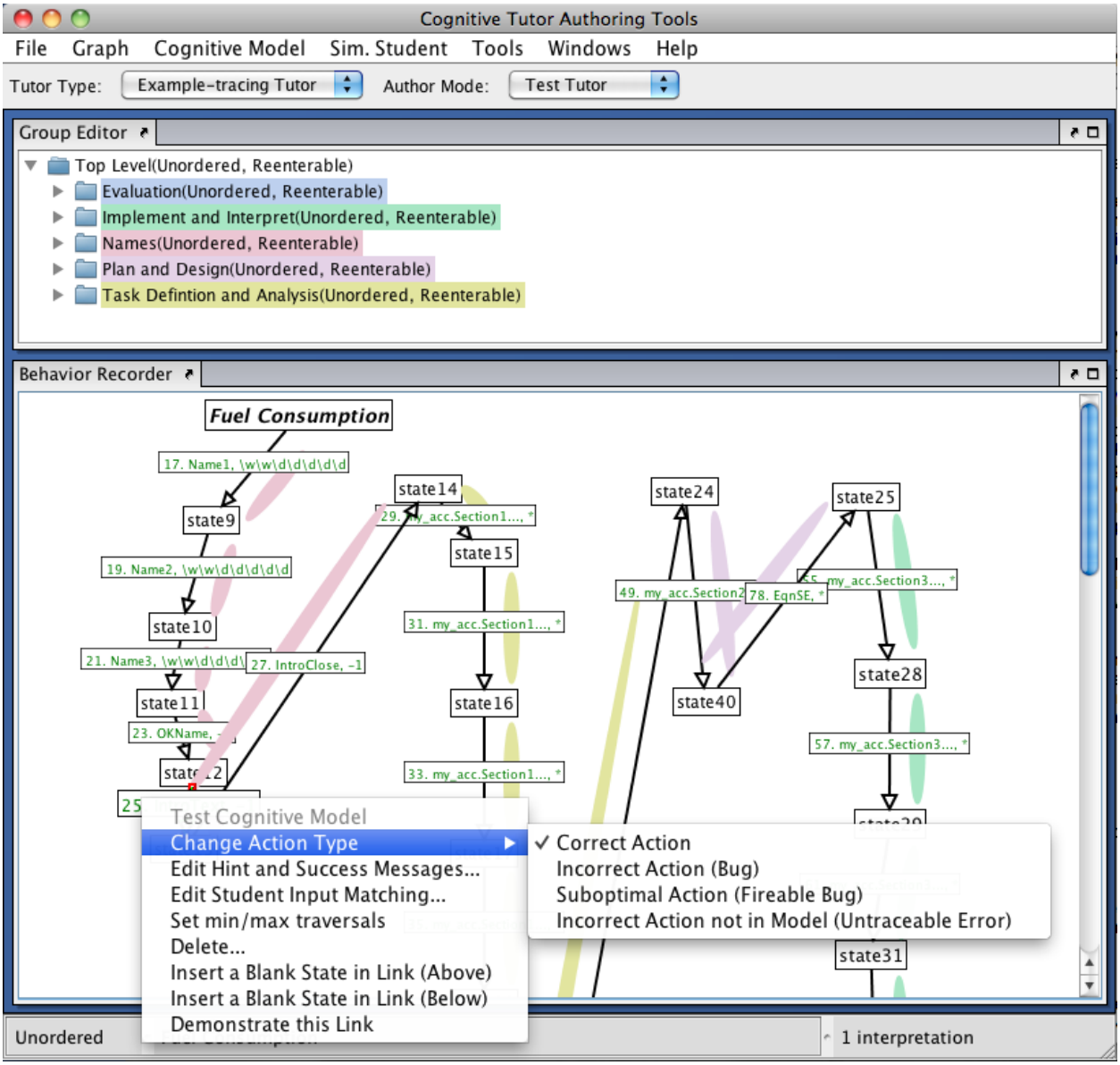


Figure 3.2: Example of a section of the BRD for the fuel consumption activity in CTAT example-tracing tutor mode. The options menu demonstrates the ability to change feedback messaging, number of traversals, or whether the action is correct or incorrect. The Group Editor window along the top allows sections of the activity to be grouped together (and coloured for visual association), so that behaviour between and within these sections can be varied.

3.2 Activity-General Features

The nature of each invention activity used in this thesis was similar enough to allow for an equivalent structure for all activities. The combination of the Adobe® Flash® CS5 and CTAT development environments allow for a single activity-general system to be built. In particular, an activity-general interface can be built in Adobe® Flash® CS5 and CTAT's mass-production feature allows easy construction of BRDs for each of the activities from a single activity-general BRD.

3.2.1 The Activity-General Interface

A single activity-general interface was built in Adobe® Flash® CS5 for each of three levels of invention scaffolding, which are described in further detail in Chapter 4. This section will use highest level of scaffolding as an example, since lower levels need only remove individual components. Adobe® Flash® CS5 come with various standard components and tools. CTAT components are installed as an extension to the Adobe® Flash® CS5 environment for use in the interface (see components list towards the right of Figure 3.3, next to the standard vertical toolbar). Individual components have various properties that can be changed, including size, position, and whether the user can make edits to the component. Special features of the ISE interface include the accordion structure, the equation editor and the use of self-explanation.

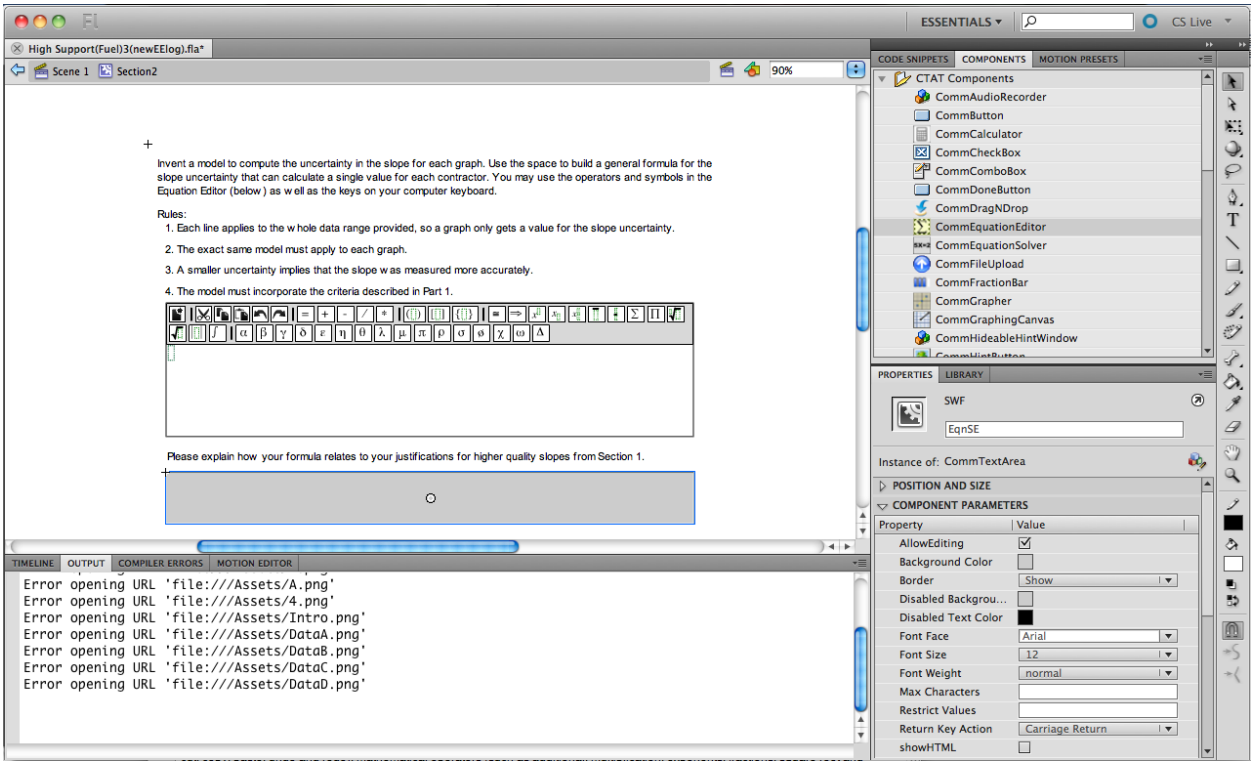


Figure 3.3: Screenshot of interface development using Adobe® Flash® Professional CS5. The CTAT components can be seen next to the standard toolbar on the right of the screen. Below the components list is a sample of the component properties, for a Text Area in this case. (Adobe® product screenshot(s) reprinted with permission from Adobe Systems Incorporated.)

The Accordion Structure

The first level of structure in the ISE was to support the cyclic and iterative nature of the activities. While this is first accomplished using the ordering and multiple traversals features of the BRD (see Section 3.1.2), the activities were all set in an accordion structure as seen in Figure 3.4. This structure permits students to easily move back and forth between sections of the activity without losing any data. Each section of the accordion holds a stage of the invention scaffolding.

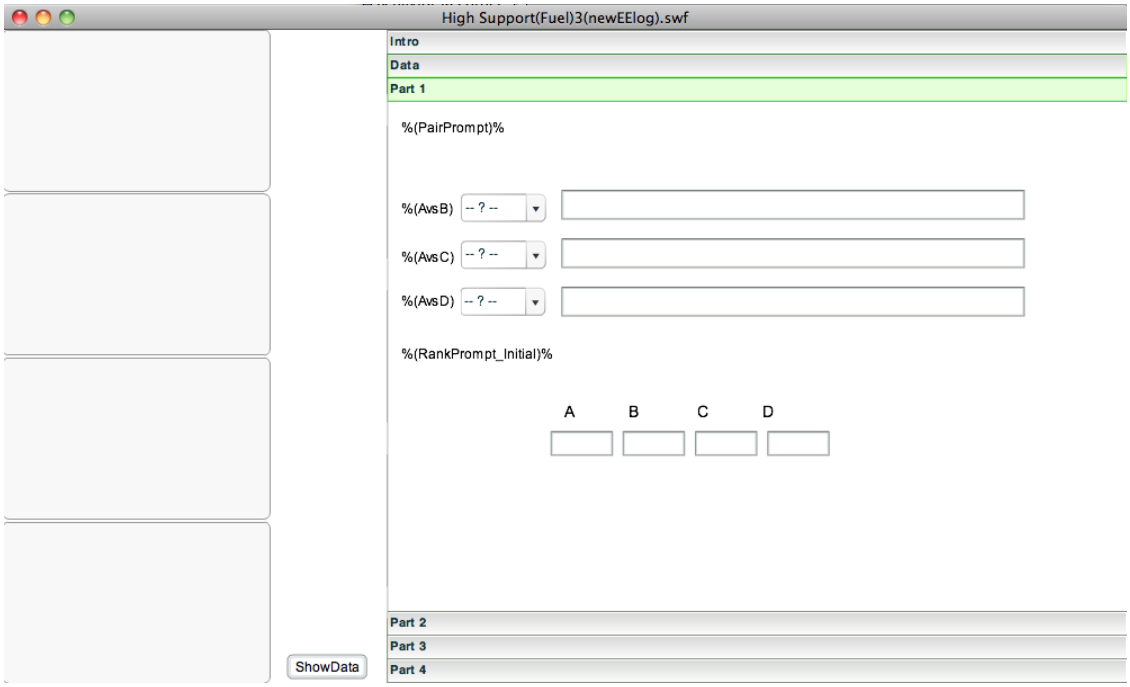


Figure 3.4: An interface skeleton for mass-production of high-level interfaces for the ISE.

The Equation Editor

For the domain formulas being used in the ISE in the first-year physics lab, it was deemed important for students to be able to develop general mathematical formulas for their inventions before implementation. In particular, these formulas would need to include various symbols and operators used in college-level mathematics. Compared to the Invention Lab [34], where students invented methods in a stepwise format where they could apply their inventions immediately to the data, students using the ISE would need to build a single general equation on the screen, as though they were designing on paper. A full Flash® equation editor was developed for this purpose by Explora Communications Inc.

This component was made up of a toolbar along the top and a text space below. The toolbar included formatting buttons (such as clear, cut, copy, paste, undo and redo), mathematical operators (such as additional, multiplication, exponents, fractions, square root and summation), different types of brackets and greek letters. The particular buttons were selected based on several criteria: basic tools that are regularly used at the level of a first-year undergraduate course, common tools and operators that were identified from handwritten invention activity solutions from previous years and additional standard tools that are used in basic mathematical formulas to generalize to future tasks. In addition, a main goal of the equation editor was to provide students with the same development opportunities as though they were creating their equations on paper. Some features that were included in order to achieve this goal were to automatically re-size the fraction lines and bracket spaces to fit their contents, and to provide visual superscripts and subscripts. Each of the toolbar buttons could be turned on or off, so that only a selection of buttons are visible to the user. Contents of the component were translated via LaTeX code in the log files. This allowed the

equations to be visibly recreated as seen by the students for analysis.

Figure 3.5 shows the equation editor with all available toolbar buttons visible (see Figure 2.2b for a sample of how the editor is embedded into the system with a subset of the available toolbar buttons visible).

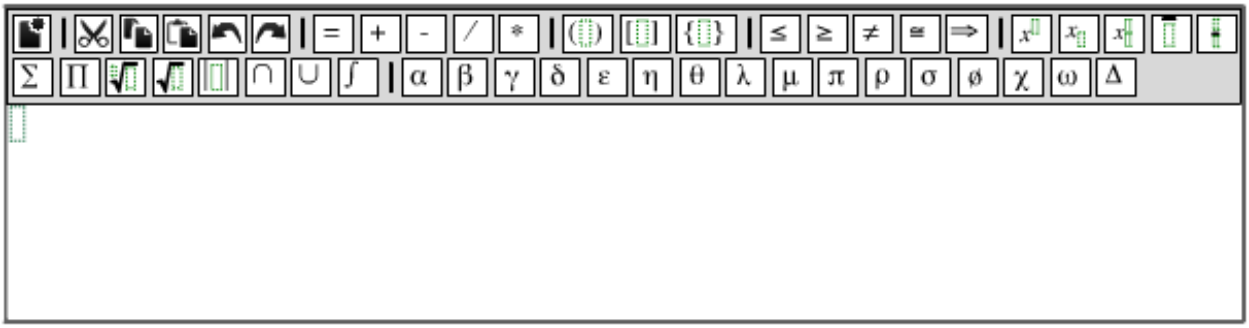


Figure 3.5: An Equation Editor was developed for the ISE so that students could design and visualize complex mathematical formulas. All available toolbar buttons are visible and include formatting options, mathematical operators, various brackets and greek letters.

Self-Explanation

In order to engage students in self-explanation, various sections of the task include prompts to justify a previous action (such as after constructing their inventions in the equation editor). These prompts would be followed by a text area where students can type their comments. The colour of the student-input areas were coloured grey in order to differentiate them from the non-editable interface prompts.

3.2.2 Mass-Production

One benefit for building the system in CTAT is the "mass production" feature that makes creating new problems very straightforward. The mass production mode uses a generalized, skeleton BRD with placeholders in each component in the interface, in place of activity-specific content (such as in Figure 3.4). The interface components in the skeleton BRD are filled with placeholders in the form of *%(ComponentLabel)%*. For example, individual images use placeholders such as *%(Plot1)%* to label the first graph, while the prompt to compare the contrasting cases is labeled *%(CompPrompt)%*.

Once the skeleton BRD is developed, a spreadsheet is generated with each component placeholder down the first column (see Figure 3.6). Each activity is added in subsequent columns and the associated activity-specific prompts corresponding to each placeholder are entered in the appropriate rows. Once the spreadsheet is completed for the necessary activities, CTAT links to the file and uses each column to generate a BRD with the activity-specific information. This is a particularly useful feature for making invention activities easy to create and manipulate, since new developers need only understand how to manipulate a spreadsheet in order to produce new activities of the same structure.

	A	B	C
1	Problem Name	Fuel Consumption	T-test
2	%(startStateNodeName)%		
3	%(Intro)%	Assets/Intro.png	Assets/Intro.png
4	%(Plot1)%	Assets/A.png	Assets/A.png
5	%(CompPrompt)%	In each of the following pairs, which of the contractors does a better job of measuring the slope of the data and why?	In each of the following pairs, which company's machines are least likely to break within 2 hours of operation?

Figure 3.6: A sample spreadsheet produced in Mass Production mode of CTAT. Each of the component placeholders are found in column A, and the corresponding activity-specific content are found in the subsequent columns (fuel consumption task is in column B and t-test activity is in column C).

3.3 Log Data

Individual actions made by a user in the interface are logged through connection with the BRD. A time-stamped log is created for each student action (such as typing into a text box or clicking on a button) once the component they were working in loses focus (for example, by selecting a new component). The log information also includes information about the user, the component in use, the action completed and the data input. In the case of the ISE, a single log file was created and appended every 24 hours and would include all information from users accessing the system during that time. Since this creates a rather long and complex file of data, the log file is converted to a spreadsheet format with individual columns for time, user information, activity name, selection, action and input. Through this spreadsheet, users' performance can be examined on a variety of levels. For example, for the analysis of the Machine Malfunction invention activity (see Chapter 6), each student's final input into the equation editor was used for the analysis of features and comments included in their inventions. In contrast, one could examine each student's series of log submissions into the equation editor to observe patterns of behaviour while developing their formulas. The log data is particularly beneficial for researchers, but may provide overwhelming amounts of information for instructors. The ISE therefore provides an online environment for classroom use, which is described in the next section.

3.4 ISE in the Classroom

The online-environment for the ISE allows individual accounts to be created for students and instructors. A student account has access to individual assigned activities and history of previous completed activities. The instructor account, on the other hand, has various class management features. An instructor can form classes and create student accounts that belong to a particular class. Individual activities can be uploaded and assigned to individual students or an entire class. The activities can be in the form of individual tasks or as a series of activities that form a problem set. In the case of invention activities, these problem sets can form the full invention process, by having the invention activity assigned as the first task, with instruction and practice activities assigned as the subsequent problems.

Furthermore, the instructor account has access to individual students' progress and can view

the entire class' progress through a problem set, including statistics such as how long they took to complete a problem set. These features provide information about student performance on the tasks, which can motivate additional activities (for example, if the student took an exceptional amount of time to complete the practice section of the activity), or inform the instructor that students understood the content and can therefore move on. This online information is much more streamlined than the log data, and can be more efficient at monitoring the class performance than examining individual student actions in the system.

3.5 Conclusions and Future Directions

The ISE has various features to improve use of invention activities in the classroom and for research-purposes. The example-tracing mode available through the CTAT development environment make generation of invention activities straightforward and requires no programming experience. CTAT also has a mass-production feature that can create multiple activities from the same interface and BRD. This makes the system easily transferrable to new tasks and facilitates development and use by new instructors. In particular, future users of the ISE need only manipulate a single spreadsheet to create new activities that fit in the research-based scaffolding. The logging features allow fine-grained analysis to be carried out for students' invention behaviours during the tasks, so that further research can be carried out to continue to evaluate the effectiveness of invention activities.

Use of the system during the study described in subsequent chapters of this thesis has motivated several changes and improvements. The current version of the system requires students to access an external spreadsheet software to implement their inventions. This means significant amounts of data are missed in the log files, since none of these actions can be tracked. This is especially problematic since it was observed that students often turn to a spreadsheet to design their formulas long before they've entered them into the equation editor. In order to develop a stronger image of students' invention processes, it would be beneficial to embed a spreadsheet tool into the system. This would allow their implementation and development actions to be logged in real-time, and supplement analysis of student behaviour and reasoning.

The current logging function of the equation editor also seems to miss significant data during student planning and design. While the general logging function, that of logging contents of components when they lose focus, is useful for most components, details of the reasoning students undergo when designing and planning their equations is lost if they continue to develop their formulas in the equation editor. That is, they may come up with multiple inventions while in the equation editor long before the contents are logged. With this in mind, we plan to implement new logging strategies for the equation editor. The editor will log after a certain time period, after a certain number of key strokes or selection of toolbar buttons, and immediately before the user clears the screen or uses a delete or backspace function. All of these items will have preferences in the interface in order to enable or disable the logging feature and to specify the length of time or number of keystrokes. These changes would allow a more detailed understanding to be developed of how students invent formulas. While this may create a significant load of data, final contents would still be accessible for other forms of analysis. The continuous stream of information about actions in the editor would be used only for study of patterns of behaviour.

Lastly, the invention activities used in the system thus far have only included the invention process, with the instruction and practice phases being led by the course instructor outside of the system. Embedding the instruction and practice activities into the system would make the tasks further generalizable to new instructors for easy implementation in new classrooms. While each activity is included in a single interface (to permit continuous movement between all areas of the task), a series of interfaces (also known as individual problems) could be combined into a problem set. A student can be assigned an individual problem set, meaning they must work sequentially through each of the problems in the series, and cannot access a new problem until they have completed the previous one. This feature is useful for inventions, since it means the invention activity interface can be made the first problem in a set, with interfaces for instruction and practice immediately following in separate problems.

Future study with the ISE should incorporate the above features.

Chapter 4

Method

This chapter describes how the ISE was used to study the effects of varying levels of scaffolding in invention activities. Section 4.1 describes the context and conditions for the study. Section 4.2 briefly describes the assessment items. Section 4.3 outlines the various statistical methods used for analysis of the assessments in the following chapters.

4.1 Design

A two-group pretest-treatment-posttest experimental design was used over a four month period. Participants were 134 students from a first-year physics laboratory course at the University of British Columbia. The course was supported by five different teaching assistants (TAs), including the author, and a course instructor. The TAs worked in pairs during each of the lab sections, and the course instructor introduced the lab activities and provided follow-up instruction after the invention activities. The preceding course used invention activities that focused primarily on graphical representations of data. The students were, therefore, familiar with invention activities, but the nature of the topics for this study were mathematical formulas to describe statistical data. The course supplements two concurrent honours courses, both of which historically attract very high-achieving students. Out of high school, these students score approximately 70% [11] on the Force Concept Inventory [23]. The first course, Physics 109, is an enriched physics course for first-year students in any science program. The second course, called Science One, is a multi-disciplinary program where all students learn Chemistry, Biology, Physics and Mathematics curriculum in a collaborative, team-taught, interdisciplinary setting. About 40% of students in Physics 109 and 10% of students in Science One continue in the second physics program [30]. Most of the remaining students plan to attend medical school, although are not tracked beyond first year.

Both groups were given five invention tasks spread throughout the 12-week course. Students were awarded participation marks for completing invention activities, therefore they were very low-risk tasks that would not affect their mark in the course. Students generally worked in groups of 2 or 3 (except during the near-transfer invention activity, where they worked individually). They were assigned seating that ensured students at each table were in the same condition, but individual pairs rotated each week. This was to ensure that students worked with different individuals each time within their conditions. Each activity was followed by direct instruction on the target domain from the course instructor, and the students would then be required to practice the formula during the analysis phase of that week's physics experiment. Students were randomly assigned, within lab sections, to one of two groups. The "Unguided Invention" (UI) group received inventions that always had low levels of scaffolding (N=73). The "Faded Guidance" (FG) group received tasks with scaffolding that faded across activities from high-level during the first two inventions,

medium-level during the subsequent two activities and finally a single unguided activity (N=74). The structure of the three scaffolding stages can be found in Table 4.1. A timeline of the study, including assessments and fading levels can be seen in Table 4.2.

<table>
<tr><th>Invention Skill</th><th>High</th><th>Medium</th><th>Low</th></tr>
<tr><td>Task Definition</td><td colspan="3">Introduction Story</td></tr>
<tr><td rowspan="2">Analysis</td><td>Explicit Pairwise comparisons with self-explanation</td><td>Overall explanation to describe features present</td><td rowspan="2"></td></tr>
<tr><td colspan="2">Qualitative ranking of all data sets</td></tr>
<tr><td rowspan="2">Plan and design</td><td colspan="3">Invent a formula</td></tr>
<tr><td>Explain the solution based on hypotheses</td><td></td><td></td></tr>
<tr><td>Implementation & interpretation</td><td colspan="3">Implement the solution for all data sets</td></tr>
<tr><td rowspan="2">Evaluation</td><td colspan="2">Rank data sets based on implementation</td><td rowspan="2"></td></tr>
<tr><td>Explain initial and final rankings</td><td></td></tr>
</table>

Table 4.1: The table describes the types of prompts used at each level of metacognitive scaffolding. The UI group consistently received low-levels of support on each activity, whereas the FG group received fading levels from high to medium to low across the study. Blank cells in the table mean the specific skill was unscaffolded.

The name and topics of the inventions and the associated target formulas were: Planet Phaedra, which addressed linear least squares fitting, Equation 4.1; The Not-so-Grand Canyon, which introduced the weighted average, Equation 4.2; Glucose Oxidation, which presented weighted linear least squares fitting, Equation 4.3; Fuel Consumption, which addressed uncertainty in the slope of a best-fit line with zero intercept, Equation 4.4; Lab Books, which presented uncertainty in the slope of a best-fit line with non-zero intercept and uncertainty in y-values, Equation 4.5. Details of the invention activities used for each of these domains can be found in Appendix B.

$$\chi^2 = \frac{1}{N}\sum_{i=1}^{N}\Big(y_i - f(x_i)\Big)^2 \tag{4.1}$$

$$\overline{x_w} = \frac{\displaystyle\sum_{i=1}^{N}\frac{x_i}{\delta x_i^2}}{\displaystyle\sum_{i=1}^{N}\frac{1}{\delta x_i^2}} \tag{4.2}$$

$$\chi^2 = \frac{1}{N}\frac{\displaystyle\sum_{i=1}^{N}\Big(\frac{y_i}{\delta y_i^2} - \frac{f(x_i)}{\delta y_i^2}\Big)^2}{\displaystyle\sum_{i=1}^{N}\frac{1}{\delta y_i^2}} \tag{4.3}$$

$$\delta m^2 = \frac{1}{N} \frac{\sum_{i=1}^{N} \Big(y_i - f(x_i) \Big)^2}{\sum_{i=1}^{N} x_i^2} \tag{4.4}$$

$$\delta m^2 = \frac{1}{N} \frac{\sum_{i=1}^{N} \Big(\frac{y_i}{\delta y_i^2} - \frac{f(x_i)}{\delta y_i^2} \Big)^2}{\sum_{i=1}^{N} \frac{x_i^2}{\delta y_i^2}} \tag{4.5}$$

A second manipulation was also applied during the study such that two lab of the four sections were given direct instruction about expert invention strategies and two were not. The instruction was provided through a short lecture at the start of the study before the students had participated in any of the invention tasks for the study. The instruction was also provided as written notes displayed on the overhead screens in the classroom throughout the third invention activity on weighted least squares fitting, to ensure that students had had enough exposure to the structure of the invention tasks to be prepared to receive the instruction [38]. This condition was applied to determine whether direct instruction about the reasoning behaviours was sufficient for students to use them without prompting, or beneficial to assist students when prompted [41]. Unfortunately, the effect of the instruction was overwhelmed by inherent differences in strengths of the lab sections, meaning invention instruction conditions were not equivalent at the outset of the study. This condition was thus removed from all analysis, and the section effects were accounted for in all analysis.

4.2 Assessments

Various assessments were used to provide insight into the acquisition of domain-level knowledge and scientific reasoning skills. The timeline of the delivery of the assessments is found in Table 4.2. Comparisons of domain-level knowledge were directly measured through a statistics diagnostic test developed by the researchers and course instructor, administered before and after the experimental treatment period. Further information about this assessment is provided in Chapter 5 and the full test can be found in Appendix A. Three questions were added to the post-test that asked students to evaluate deliberately flawed variations of equations presented in the invention activities throughout the term. This assessment is described in Section 5.2.

A transfer invention task, that is a task set in a new domain where students are required to transfer the skills they have learned, was created to measure differences across groups in scientific reasoning behaviours as well as domain-level performance. Since all invention activities throughout the study focused on statistics and data analysis processes, the near-domain transfer task asked students to invent the independent one-sample t-test statistic. This activity and the analysis of student inventions is detailed in Chapter 6. One week after this task, students were asked to redraw the data and canonical solution to the activity. The details of this assessment can be found in Chapter 7.

Week	Invention Activity Title	Invention Activity Domain	FG Scaffolding	Assessments
1				Statistics Diagnostic Pre-Test
2	Planet Phaedra	Linear least-squares fitting	High	
3				
4	Not-so-Grand Canyon	Weighted Average	High	
5				
6	Glucose Oxidation	Weighted Linear Least Squares	Medium	
7				
8	Fuel Consumption	Slope Uncertainty with zero intercept	Medium	
9				
10	Lab Books	Slope Uncertainty with δy and non-zero intercept	Low	
11	Machine Malfunction	T-value	Low (completed individually)	Statistics Diagnostic Post-Test
12				Recreate Data

Table 4.2: Timeline of invention activities, fading levels of support for the FG group and assessments delivered throughout the study.

4.3 Analysis Methods

Background knowledge was assessed with the Concise Data Processing Assessment [17], a diagnostic which probes students' understanding of uncertainty and fitting models to data. Assessment items were compared between groups using one of two analysis methods. Analysis of covariance (ANCOVA) was used for continuous data with scaffolding condition as a factor, and CDPA scores and lab section as covariates. Significant results are described with the F-value and associated degrees of freedom, and the p-value for the test. For results in the form of binary data, logistical regression models (or logit models) were used to examine effects of scaffolding condition, controlling for CDPA scores and lab section. Significant results are described with the test's Z-score and p-value.

Lab section was included to account for additional environmental factors that may affect classroom performance. The instruction condition that was initially included in the study was found to be confounded by effects of individual lab sections, even when controlling for CDPA scores. In particular, one of the four lab sections, held on a Monday afternoon, significantly outperformed others on several measures, while a second lab section, held on a Friday afternoon, significantly underperformed compared to other sections. These differences may be due to selection bias with regards to which students chose these lab sections, or environmental (or perhaps social) factors with regards to academic performance on a Monday afternoon compared to late on a Friday afternoon.

In text and tables, binary results are presented as average counts. Results in the form of continuous data are presented as an average with the standard error following parenthetically, such as mean (stderr). Significant differences are described using p-values, which describe the probability that the observed differences are due to chance. Statistically significant results, those with p values less than 0.05 are highlighted in tables using superscript symbols that refer to particular ranges of significance: † refers to a marginally significant results with a p-value less than 0.1 but greater than 0.05; * refers to a p value that is less than 0.05 and greater than 0.01; ** corresponds to a p-value less than 0.01 and greater than 0.001; *** corresponds to p-values less than 0.001.

Chapter 5

Assessment: Statistics Diagnostic

This chapter describes the statistics diagnostic used to assess learning of conceptual and procedural domain topics. Section 5.1 describes the items used on both the pre- and post-tests. Section 5.2 describes debugging items that were unique to the post-test. Sections 5.3 and 5.4 present the results and conclusions of the assessment, respectively.

5.1 Pre- and Post-Test Items

A multiple choice statistics diagnostic was created for two purposes. Firstly to confirm group equivalence at the start of the activity through a pre-test. Secondly to measure learning of the domains presented in the study.

The test items, which can be found in Appendix A, were developed by the researcher and course-instructor. The individual questions were chosen to cover concepts from three of the study's invention activities that were most well-developed by the start of the study (namely linear least squares fitting, slope uncertainty with an intercept of zero, and weighted average). The test posed three procedural and two conceptual questions in order to directly evaluate how well the invention process helped students learn the content. This was in contrast to the other assessment items described in later sections that aim to measure various reasoning abilities.

The pre-test was administered during the first week of the term to measure the students' abilities at the outset of the study, especially to compare initial equivalence of experimental groups. The post-test was administered during week 11 of the study, after each of the five main invention activities had been completed. In both cases, students were not informed about the diagnostic in advance, so students would not have studied for the test. On the day of the assessment, students were told that their scores on the test would not hurt their course grade, but could contribute positively if they did well.

5.2 Post-Test Debugging Items

Three 'debugging' questions were added to the end of the statistics diagnostic post-test to measure conceptual understanding of the technical features of the equations presented during the study. These items presented students with predesigned variations on formulas previously presented during the study with deliberate flaws. Namely, the questions related to linear least squares fitting, weighted average and slope uncertainty with an intercept of zero. Students were asked to identify whether the equations were valid and to justify their answers.

It has been shown that while evaluating methods is not sufficient replacement for invention activities with respect to learning outcomes, invention does improve performance on debugging

items [35]. It is hypothesized that students who better understand the features of the domain (presumably the FG group) would be better at debugging the formula variations.

For each question, students' justifications were assessed based on features they discussed and whether they recognized the technical qualities required from the equations to support the goals of the item. Analysis for these items was carried out by two raters (one of whom was the author), with an average correlation greater than 90% across all items.

5.2.1 Linear Least Squares Fitting

$$\chi^2 = \frac{1}{N}\Big(\sum_{i=1}^{N}\big(y_i - f(x_i)\big)\Big)^2 \tag{5.1}$$

The first item, Equation 5.1 is a variation on the linear least squares residuals formula (Equation 4.1 demonstrates the correct formula), with the exponent outside of the summation, rather than inside. This technical change would not account for negative values in the sum, meaning large negative and positive differences could cancel. This could give a small value of χ^2 for very bad fits. Students received credit if they noticed the misplacement of the exponent, and commented on the issues it creates.

5.2.2 Weighted Average

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i \delta x_i}{\sum_{i=1}^{N} \delta x_i} \tag{5.2}$$

The second item, Equation 5.2, is a variation on the formula for weighted averages. This item varies the way the uncertainties in the data points, δx_i, are weighted with respect to the data points, x_i. That is, Equation 5.2 provides more weight to data points with larger uncertainties. The accepted formula, Equation 4.2, correctly weights values with lower uncertainty by multiply each data point by the inverse of its uncertainty. Credit was therefore given to responses that recognized that the formula was not valid because it added more weight to items with larger uncertainties. This would assess whether students not only understood the goal of the formula, but also could connect that goal to technical components of the formula.

5.2.3 Slope Uncertainty with Zero Intercept

$$\delta m^2 = \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{y_i}{x_i} - m\Big)^2 \tag{5.3}$$

Equation 5.3 is a variation on a common student invention for the slope uncertainty with zero intercept activity seen in past uses of the activity[36]. This solution looks at the residual differences between the slope of the best fitting line and the slope of lines between individual points and the origin (the correct form of the formula is seen in Equation 4.4). The authors found that this formula was quite valid at addressing the key features of the domain (vertical residuals, leverage and sample size), and therefore accepted answers with justifications that touched on these items. One key area

where the formula is not valid is for data points very near to the origin (that is, $x_i \approx 0$), since this would cause a value in the sum to approach infinity. Responses that determined the formula was not valid for this reason were therefore also credited.

5.3 Results

ANCOVAs were used to determine any statistically significant differences across conditions, controlling for background knowledge through the use of the CDPA and lab section, described in Chapter 4.

Pre-test data on the statistics diagnostic demonstrated that students in both conditions were initially equivalent with regards to domain-level knowledge (see Table 5.1). While differences between conditions on the post-test were not statistically significant, overall there was significant pre- to post-test improvement: Pretest = 2.15(0.10), Posttest = 2.71(0.08); F(3,126) = 12.3; p = 0.0006. The ANCOVA across conditions for post-test also demonstrated that the CDPA is a significant indicator of ability, with differences in statistics diagnostic post-test scores aligning with differences in CDPA scores: F(3,126) = 8.6; p = 0.004.

Item type	Test Time	Unguided Invention	Faded Guidance
Conceptual (/2)	Pre	1.31 (0.08)	1.17 (0.09)
	Post	1.51 (0.06)	1.48 (0.08)
Procedural(/3)	Pre	0.93 (0.10)	0.89 (0.10)
	Post	1.25 (0.08)	1.18 (0.08)
Overall(/5)	Pre	2.24 (0.15)	2.06 (0.13)
	Post	2.76 (0.11)	2.66 (0.11)
Debugging (/3)	Post	0.76 (0.10)	0.98 (0.10)*
$\dagger - p < .1; * - p < .05; ** - p < .01; *** - p < .001$			

Table 5.1: Pre- and post-test statistics diagnostic scores on procedural, conceptual and debugging items. Differences between conditions on the debugging items were found to be statistically significant.

Overall, performance on the debugging items was quite low (see Table 5.2), with an average score of 0.73 ± 0.07 out of 3 (that is, 24%). In particular, performance on the least squares debugging item was much lower than on the other items (mean = 0.14 across conditions). There was a significant difference across scaffold conditions for the overall score on the weighted average item: UI = 0.28; FG = 0.40; Z = 1.9; p = 0.06). The overall debugging score also showed significant differences: UI = 0.76(0.10); FG = 0.98(0.10); F(3,126) = 3.8; p = 0.05. The FG group outperformed the UI group in both cases.

Item	Unguided Invention	Faded Guidance
Linear Least Squares	0.16	0.26
Weighted Average	0.28	$0.40^{\dagger}$
Slope Uncertainty	0.31	0.32
$\dagger - p < .1; * - p < .05; * * - p < .01; * * * - p < .001$		

Table 5.2: Statistics diagnostic debugging item scores by topic. Differences between conditions on the weighted average item were found to be statistically significant.

5.4 Conclusions

A five-question statistics diagnostic was developed to assess students' domain-level knowledge, before and after experimental treatments. The statistics diagnostic was made up of two conceptual and three procedural questions relating to linear least squares fitting, weighted average and slope uncertainty with an intercept of zero. No statistically significant differences were observed between scaffolding conditions, but 11% learning gains were observed across the study. FG students demonstrated significantly stronger performance on the debugging items, suggesting three explanations.

Firstly, if the FG students had acquired better domain knowledge through the scaffolded inventions, then they would have better understood the components of the formulas, and thus noticed when they were missing. It is understandable that the groups would perform equivalently on the conceptual or procedural items of the diagnostic, since these items examine how or when to implement the formulas, which is not addressed in the invention activity. In contrast, the debugging items required deep understanding of how the technical features of the formulas addressed the conceptual purpose of the equations, which is supported by full use of the invention process. This explanation is supported by various studies that demonstrate how invention activities improve domain knowledge and future learning [35] [39].

A second explanation is that the FG group may have improved photographic memory of the original formula. This explanation, however, would not justify why their explanations were improved. That is, memory of the formula would improve a student's ability to recognize what about the formula was incorrect, but would not affect their ability to explain why it is invalid.

A final possibility is that the students in the FG group developed better use of monitoring and evaluation skills than the UI group. Their improved performance on these items would therefore reflect improved ability to evaluate the formulas and explain their limitations. This explanation is also supported by studies that demonstrated that domain-general scaffolding improved performance on self-checking, monitoring and evaluation items [10].

The original explanation is supported by further examination of performance on individual debugging items. In particular, the FG group appears to outperform the UI group on the two debugging items associated with invention activities where the FG group received the highest level of scaffolding (Planet Phaedra, and the Not-so-Grand Canyon). Once the FG group's scaffolding was faded to medium (on the Fuel Consumption activity which is associated with the third debugging item), the two groups were equivalent in their ability to debug the formula. This suggests that the use of explicit prompts to make pairwise comparisons, used only in the high-level scaffolding,

improve understanding of the technical features of the domain. The apparent differences in the first debugging task, however, were not found to be statistically significant when controlling for background knowledge and section. In which case, this observation may only describe the difficulty of the individual items.

While it is understood that invention activities improve domain knowledge compared to traditional instruction [33] [39], this assessment confirms that scaffolding of reasoning behaviours adds further improvement to deep understanding of the functional components of the domain equations [35]. An additional result from the analysis is that students in the FG group demonstrated better performance on the debugging items more than two months after learning the associated domains. This demonstrates that the students were able to retain the domain knowledge after significant time delays.

Chapter 6

Assessment: T-test Invention Activity

This chapter describes the Machine Malfunction invention activity, a near-domain transfer task used as an assessment of invention skills. Section 6.1 introduces the domain and the activity assigned. Section 6.2 outlines the various analysis methods used. Section 6.3 describes the results obtained from the analysis. Section 6.4 summarizes the activity and draws conclusions based on the results obtained.

6.1 Introduction

Since all invention activities throughout the study focused on statistics and data analysis processes (such as uncertainty in best-fit line parameters, linear least-squares fitting, or weighted means), a near-domain transfer invention task had students invent the independent one-sample t-test statistic seen in Equation 6.1:

$$t = \frac{\bar{x} - \mu_o}{\frac{\sigma}{\sqrt{N}}}, \tag{6.1}$$

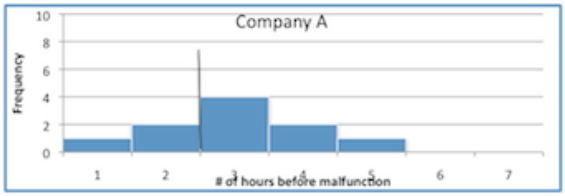
where $\bar{x}$ is the sample mean of the data set, μ_o is the reference value that the sample mean is being compared to, σ is the standard deviation of the data set and N is the sample size.

The task presented a story about a computer company whose machines were malfunctioning after only two hours of operation. The students' task was to create an index to help the company evaluate which of five computer suppliers' machines was least likely to malfunction before two hours of operation. To do this, students were provided with data from the five suppliers, showing the frequency of hours of machine operation before breakdown (see Figure 6.1).

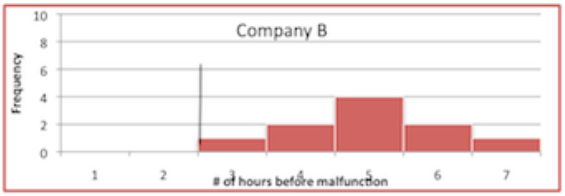
While both groups received low scaffolding during the task, meaning they were not explicitly asked to make pairwise comparisons, examination of the contrasts between Company B and the other graphs would extract three domain features. First, the further the mean of the data set is from the two-hour mark, the more reliable the company's machines. Therefore, Company A's machines are more likely to break down than those of Company B due to the position of the mean of the histogram. Mathematically, this is represented as the distance between the population mean, $\bar{x}$, and the reference value, μ_o, in the numerator of Equation 6.1. In this case, $\mu_o = 2$. Secondly, the more data collected, the more reliable the determination of the mean. Therefore, since Company C's data set includes twice as many data points as Company B, the standard error is reduced, giving it a better index. This is mathematically represented as a relation to N in Equation 6.1. Finally, the more narrow the distribution of data, the more predictable the result of future tests, making Company D's machines more reliable than those of Company B. The t-value statistic represents this as a division by the standard distribution of the data, σ, in Equation 6.1. Company E's data set presents a sample with very few data points, a large distribution and a mean shifted far from

the two-hour mark. This confounded data set was included with the contrasting cases to address issues of point versus set reasoning [9]. It aimed to encourage students to generalize the features to new cases. It also aimed to confront students with the issue that a large mean does not necessarily qualify as a better measurement, thus encouraging students to interact more deeply with the other contrasts.

Comparisons between any of the other items demonstrate confounded changes (for example between Company A and Company C, both the mean and the number of samples is changed), making it challenging to qualitatively determine which is least likely to break down before two hours.

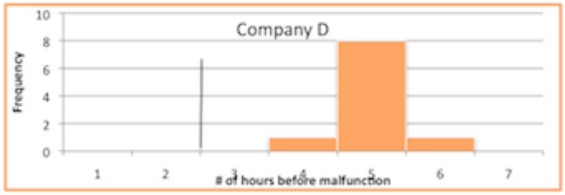


(a) Data from Company A with $\sigma_A = \sigma_B, N_A = N_B$, and $\overline{x_A} < \overline{x_B}$



(b) Data from Company B with $\overline{x_B}$, σ_B, and N as references to extract the domain features.

(c) Data from Company C $\overline{x_C} = \overline{x_B}, \sigma_C = \sigma_B$, and $N_C = 2N_B$



(d) Data from Company D with $\overline{x_D} = \overline{x_B}, N_D = N_B$, and $\sigma_D < \sigma_B$

(e) Data from Company E with $\overline{x_E} > \overline{x_B}, N_E < N_B$, and $\sigma_D > \sigma_B$

Figure 6.1: The data for the machine malfunction invention activity. The five graphs represent the frequency of machines' operating hours until breakdown from five computer suppliers. Pairwise comparisons between (b) and each of (a), (c) and (d) demonstrates that a higher index corresponds to a data set with an average shifted far away from two hours, more measurements of computers, and a more narrow distribution, respectively. These comparisons represent features for inventing the t-value for the independent one-sample t-test.

6.2 Invention Analysis

Student inventions were qualitatively analyzed using an already developed rubric with slight modifications [36]. Inventions from the ISE were converted into a PDF document using the LaTeX translation from the equation editor logs and merged with student comments in the unprompted

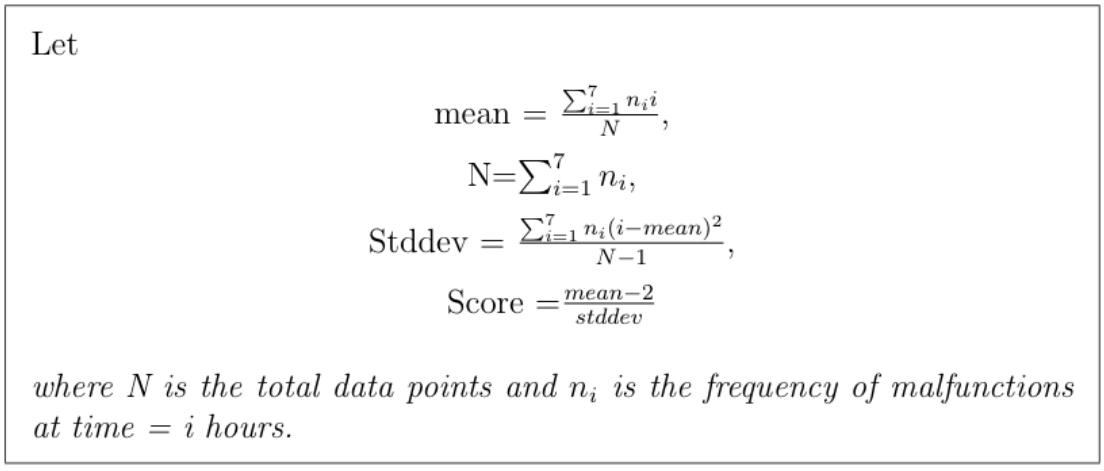
self-explanation space below the editor (see Section 2.1.3). While students may have produced multiple equations, only the last version of their formulas and comments were used in the analysis, in order to maximize the number of concepts developed by the students included in the analysis. Inventions were analyzed for features, type and focus of self-explanations, and quantitative rankings.

6.2.1 Features Analysis

Equations and self-explanations were analyzed for the intentional presence of features of the data described above. For each feature, students received a score of: 3 if the feature was present and in the correct representation (for example, subtraction of the mean from the reference value); 2, if the feature was present and operating in the right direction (for example, the mean in the numerator); 1, if the feature was present but operating in the wrong direction (for example, division by the mean); 0, if the feature was not present at all. Features were described as:

- Explicit instance of central tendency, unconfounded from an instance of spread and from sample size (specifically as $\overline{x} - \mu$).
- Explicit instance of spread, unconfounded from an instance of central tendency and from sample size (specifically through division by σ).
- Explicit instance of sample size, unconfounded from an instance of central tendency and from spread (specifically through multiplication by $\sqrt{N}$).

For example, Figure 6.2 would obtain a score of 3 for the presence of central tendency due to the correct representation of *mean - 2*. It would also receive a score of 3 for the presence of spread, correctly represented in the denominator of the score. This invention would receive a score of 0 for sample size, since it is not present in an unconfounded representation. That is, although N is present in the invention, it is only seen as part of the definitions for mean and standard deviation, not as an explicit feature.


Let

$$\text{mean} = \frac{\sum_{i=1}^{7} n_i i}{N},$$

$$\text{N}=\sum_{i=1}^{7} n_i,$$

$$\text{Stddev} = \frac{\sum_{i=1}^{7} n_i (i-mean)^2}{N-1},$$

$$\text{Score} = \frac{mean-2}{stddev}$$

where N is the total data points and n_i is the frequency of malfunctions at time = i hours.

Figure 6.2: Sample invention by a student for the Machine Malfunction activity, demonstrating the presence of central tendency and spread, as well as low-level comments.

In contrast, Figure 6.3 would obtain a score of 2 for the representation of central tendency, since the formula rewards data with a higher average hour before first breakdown but does not look at how far it is from the reference point, 2. It would obtain a score of 0 for both spread and for sample size, since they are absent as unconfounded features in the formula.

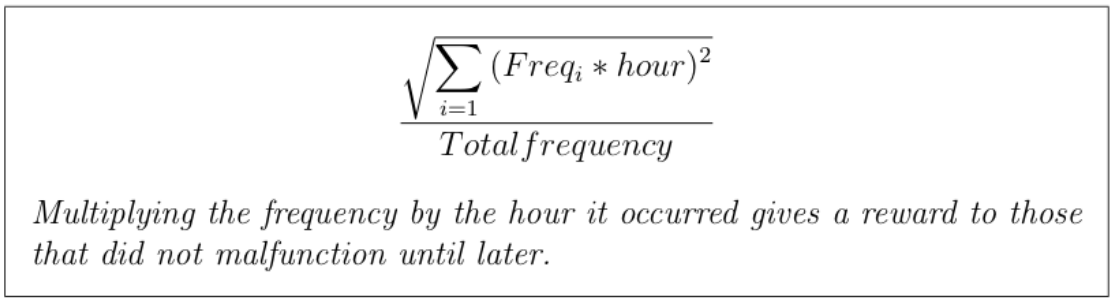


Figure 6.3: Sample invention by a student for the Machine Malfunction activity, demonstrating the presence of central tendency and a high-level comment that focuses on explaining the use of central tendency.

While this initial rubric provided very fine-grained details of the students' inventions, the absence of a feature and its incorrect representation were found to both conclude that the student had not understood or noticed the feature. These two scores were therefore reduced to a single score of 0 in a second marking rubric. Similarly, and since it is known that metacognitive scaffolding does not improve technical quality of student inventions [36], the presence of a feature in the correct direction or representation were collapsed to a single score of 1 in the second rubric, describing the presence of features. That is, the new rubric examines whether the conceptual features were present or not (for example, higher means give better scores, regardless of the distance from the reference point). Using this method, Invention 1, above, would obtain 1 mark each for central tendency and spread, but 0 for sample size, yielding a feature score of 2 out of 3. Invention 2 would obtain a feature score of 1 out 3 for the presence of central tendency.

6.2.2 Self-Explanation Analysis

Students' comments were analyzed on two levels. Firstly, students received a mark for low-level comments such as definitions or descriptions of their equations in words. They could also receive a mark for the presence of any high-level comments, such as evaluations, explanations or setting goals in terms of the features of the data. High-level comments were also analyzed to examine the focus of the comments, either on the features of the data (central tendency, spread, or sample size) or surface features (units of the equation, or magnitude of the values obtained from implementation). For example, the comment at the end of Figure 6.3 is a high-level comment as it explains why the technical format of their model takes into account the central tendency feature. This student would therefore receive 0 for the absence of any low-level comments, 1 for the presence of a high-level comment, and 1 for focusing on central tendency. In contrast, the comment at the end of Figure 6.2 and the three formulas above the Score represent low-level comments that only define terms used in their final formula for the score. It should be kept in mind that, while students in the

FG condition had been prompted to explain their inventions earlier in the study, this activity in particular was delivered with low-level scaffolding, which had no explicit prompt to include any written explanation for equations.

6.2.3 Ranking Analysis

Using students' final values from the implementation stage, quantitative pairwise comparisons were made to observe any differences in student ability to convert their qualitative understandings to mathematical representations. Students were given 1 mark for ranking central tendency if their implemented value for Graph A was smaller than that for Graph B. Similarly, a mark was award each for sample size and spread if the value for Graph B was less than Graph C and Graph D, respectively. A final ranking score out of three was given to each student (one available score for each comparison).

6.3 Results

ANCOVAs or Logit Regression Models were used to determine any statistically significant differences across conditions, controlling for background knowledge through the use of the CDPA and lab section, as described in Chapter 3.

6.3.1 Features Analysis

Overall, students noticed 1.28 ± 0.08 out of 3 possible features when creating their inventions and obtained an average score of 3.26 ± 0.16 out of 9 available points (3 for each feature). No significant difference between groups was found. For a breakdown of all values, see Table 6.1 and Table 6.2.

Feature Score	Unguided Invention	Faded Guidance
Central Tendency (/3)	1.71 (0.08)	1.89 (0.07)
Spread (/3)	0.63 (0.13)	0.69 (0.14)
Sample Size (/3)	0.76 (0.11)	0.87 (0.12)
Overall Score (/9)	3.10 (0.21)	3.45 (0.24)

Table 6.1: Machine malfunction invention activity scores for each analysis item. The items are scored out of the number in brackets in the first column of the table. Scores are given as averages with the standard deviation in parenthesis. No significant difference between groups exists.

6.3.2 Self-Explanation Analysis

In general, there was no difference between conditions on the types of self-explanations included, with approximately 65% of students making high-level comments and 30% of students including low-level comments (see Table 6.3). The distribution of comments focused mostly on the domain features (as seen in Figure 6.4) with less than 10% of students focusing on topics such as units of the formula or magnitude of the numbers that the formula produced.

Noticed Features	Unguided Invention	Faded Guidance
Central Tendency	0.74	0.87
Spread	0.20	0.24
Sample Size	0.26	0.27
Overall Features (/3)	1.20 (0.11)	1.38 (0.11)

Table 6.2: Machine malfunction invention activity binary results for whether students noticed features. Scores are given as averages with the standard deviation in parentheses, where appropriate. The FG group included central tendency in their inventions more often than the UI group.

Comment Type	Unguided Invention	Faded Guidance
High-level	67.1%	61.8%
Low-level	30.0%	29.1%

Table 6.3: Summary of levels of student comments. No significant difference was observed between groups.

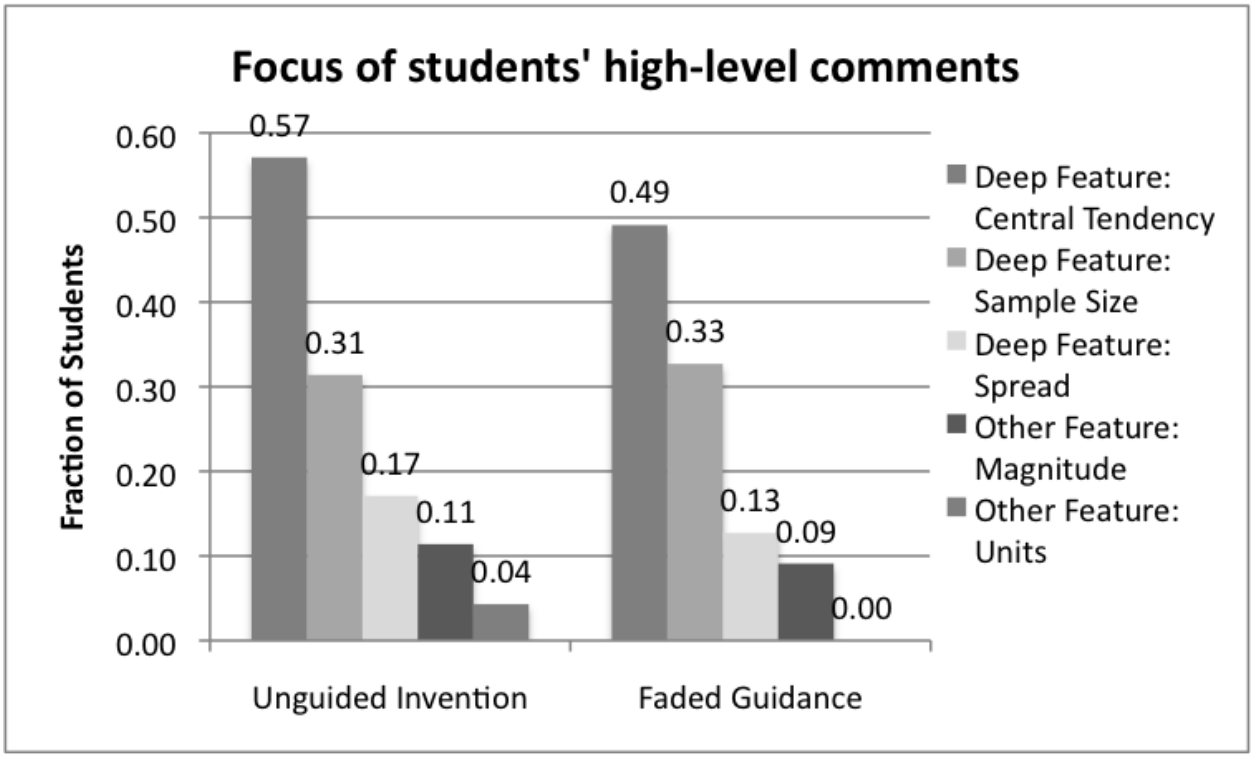


Figure 6.4: Summary of the focus of students' high-level comments. No significant difference was observed between groups.

6.3.3 Ranking Analysis

As seen in Table 6.4, there was no significant difference across conditions for ranking of graphs, based on student implementations of their inventions. Overall, these quantitative ranking scores agreed with trends in qualitative analysis of features in that most students implemented central

tendency, but there was less success with spread and sample size.

Ranking feature	Unguided Invention	Faded Guidance
Central Tendency	0.87	0.87
Spread	0.46	0.56
Sample Size	0.46	0.47
Overall (/3)	1.79 (0.11)	1.91 (0.12)

Table 6.4: Summary of quantitative rankings by associated feature. No significant difference was found between groups.

6.4 Conclusions

A near-transfer invention task was delivered to all students after completing the five main invention tasks found in Appendix B. By this point, the FG group had reached low-level scaffolding of tasks. The activity asked students to invent an index that was equivalent to the independent one-sample t-value. It was delivered as a low-scaffold task and students completed the activity individually. Student inventions were analyzed for the presence of domain features, the level and focus of unprompted self-explanations, and the ability of inventions to rank the graphs according to the domain features.

It was found that both groups performed equally well on the quality of their inventions. This implies that the scaffolding improves inventions only when directly present [36]. It also suggests that the process of fading the scaffolding was insufficient for students to internalize the invention strategies.

It is also interesting to note the lack of differences between groups in the quality of their self-explanations. This may be a consequence of implementing invention activities through the ISE. For example, the equation editor and text space may have discouraged students from freely adding comments and notes beside their formulas, as they were seen to do on paper inventions.

However, this result, combined with the lack of differences in the quality of student inventions, suggests the metacognitive scaffolding must be present to improve students' self-explanations. Since it was previously observed that the metacognitive scaffolding improves the level and focus of students' unprompted self-explanations [36], this result further suggests that the faded use of the scaffolding did not allow students to encode the reasoning abilities provided by the scaffolding. These two possible conclusions could be made clear by examination of level and focus of students' self-explanations in earlier inventions when scaffolding was present in the FG group inventions.

While no differences were found between groups on the quantitative rankings, more students were found to correctly rank spread and sample size than were found to have noticed either of these features in their equations and self-explanations. This suggests that the ranking analysis is not representative of features noticed in student inventions, especially since it was previously observed that scaffolding invention activities does not improve technically proficiency [36].

An alternate explanation to the results of this assessment is that the performance on this activity was hindered because students worked individually rather than in pairs, removing peer interaction.

Perhaps the feedback and assistance obtained from peer interaction critically supports the successful invention behaviours, and individual metacognitive skills are not sufficiently developed at this point in students' careers. This explanation is consistent with the low scores for all groups when noticing spread and sample size on the inventions. In addition, the machine malfunction task itself may have been more complicated than originally supposed, due to an expert blind spot. Students may have been overwhelmed by the representation of the data sets as histograms, and were unable to extract or implement features, such as the standard deviation, to include in their inventions.

Chapter 7

Assessment: Reproduce Data

This chapter presents a second assessment item, that of recreating data presented in the Machine Malfunction activity. Section 7.1 describes the motivation for the assessment. Section 7.2 details the methods for analyzing student responses. Sections 7.3 and 7.4 describe the results obtained from the analysis and draws conclusions about the behaviour of the two groups.

7.1 Activity Description

One week after the machine malfunction invention activity, as described in Chapter 6, students were given an activity on paper to recreate the data and solution from the t-test statistic activity. The activity briefly reminded the students of the goal of the original invention activity and asked them to redraw the data from the task to the best of their ability. The data could be reproduced in any form and students worked individually. The activity also reminded the students that the course instructor had explained "a canonical solution for this problem, called a t-value" following the invention activity, and asked the students to recreate this formula to the best of their ability.

The purpose of this activity was to provide insight into which features of the invention activity students valued as most important based on what they were able to recall. For example, whether students focused on deep features of the domain, such as the various contrasts presented in the activity, or simply remembered surface features such as matching the type or number of graphs used. This follows the work by Chase and Simon (1973) [13] whereby more expert chess players were better able to reconstruct a chess position after briefly viewing it than novice players. This was concluded to be due to the experts' improved ability to encode 'chunked' information. It is hypothesized, therefore, that FG students will have encoded each of the features of the domain, as described in Chapter 6, better than the UI students. This would be reflected in an improved ability to reconstruct the data, since the data was designed to specifically reflect those features.

7.2 Analysis Methods

Approximately 20% of the items were first evaluated independently by 4 individual raters (including the author). With an average correlation greater than 80% across all items, it was deemed acceptable for the remaining items to be analyzed by a sole rater (the author).

7.2.1 Data Coding

Students' recreations of the data were scored on two levels:

1. Whether individual domain-features were present in the data. Each of the features had to be presented with some apparent form of intention, but could be present as confounded contrasts between graphs. For example, a data set with a clearly distinct mean from another graph, but with a slightly varied distribution would represent the presence of central tendency. Four marks were therefore available in this category, one for each of the three features and an additional mark for the confounded graph (see Figure 6.1e).

2. Whether individual domain-features were present in the data as unconfounded contrasts. Since part of the intrinsic support of invention activities was to present the data in the form of contrasting cases (such that pairwise comparisons between data sets highlighted individual features) students received a mark for recognizing the importance of the contrasts. Students were given credit for features represented as unconfounded changes between graphs, that is if two graphs were equivalent except for the single modified feature. Three marks were therefore available in this category, one for each of the three possible contrasting cases representing the domain features.

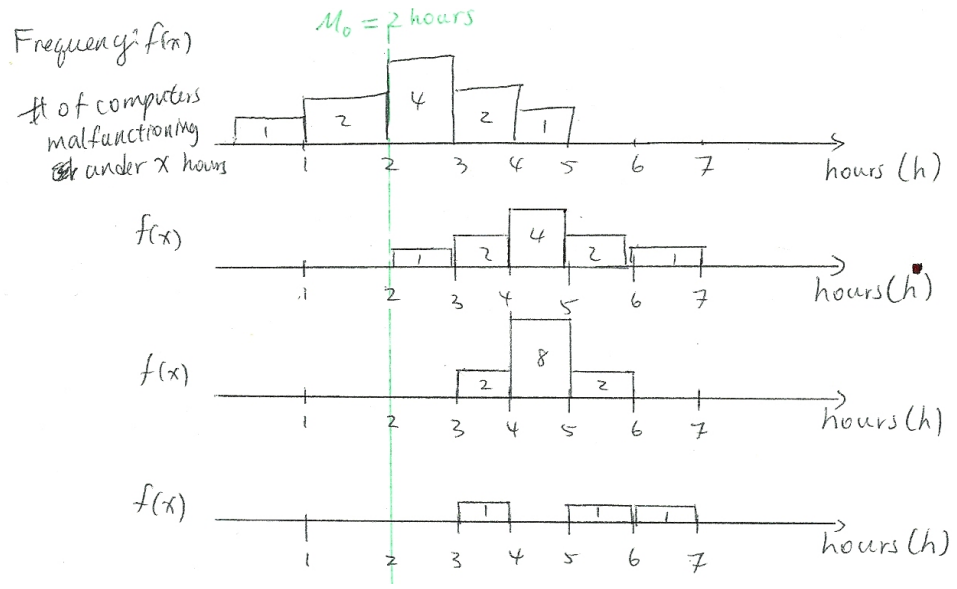
These two analysis categories, therefore, probed conceptual understanding and invention skills separately. That is, while being able to portray individual features in the reproduced graphs demonstrates understanding of the concept, displaying these features via unconfounded pairwise comparisons suggests familiarity with the invention process and skills necessary to invent.

The solution in Figure 7.1a, for example, would score a point each for reproducing central tendency (since the first graph is shifted to the right to obtain the second graph), spread (since the third graph has the same mean and sample size as the second graph, but with a more narrow distribution) and the confounded graph (final graph). They would also score a point each for reproducing central tendency and spread as pairwise comparisons. This would give them a score of 3 for recreated features and 2 for pairwise comparisons. The student who produced Figure 7.1b recreated the central tendency (an apparent shift in the mean between graph A to graph C, with spread and sample size constant) and spread (graph D shows a more narrow distribution than graph E, with the mean and sample size constant) features in a pairwise fashion. The student also recalled sample size, but not in as a pairwise comparison (graph B has a larger sample size than graphs A and C, but the means and distributions are not held constant). This student therefore is assigned a score of 3 for recreated features and a score of 2 for pairwise comparisons.

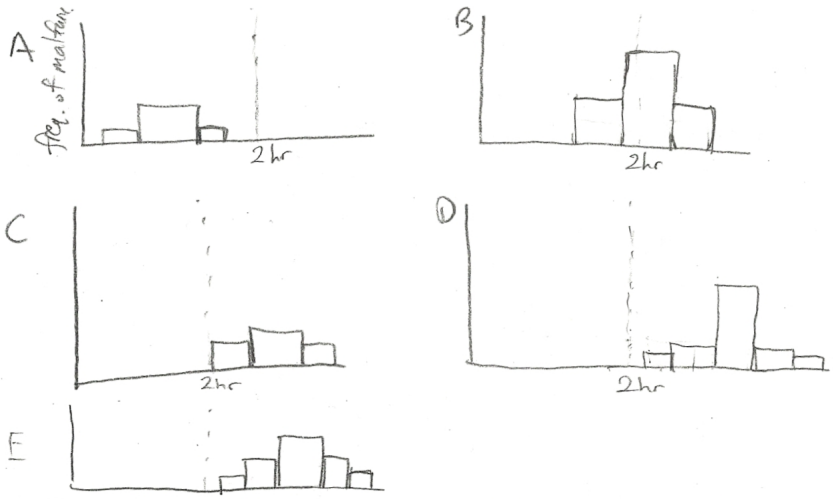
7.2.2 Equation Coding

Students' recreated formulas were coded as follows. For each feature: a score of 3 meant the feature was present and in the correct representation. That is, for the presence of $x - \mu_o$, $\frac{1}{\sigma}$, and $\sqrt{N}$; a score of 2 meant the feature was present in the correct direction (or the correct sign of the power). That is, for the presence of $\bar{x}$ in the numerator, σ in the denominator, and N in the numerator; a score of 1 meant the feature was present, but not in the right direction. That is, one mark each for the presence of $\bar{x}$, σ, and N regardless of how they appeared in the formula; a score of 0 meant the feature was absent.

Nine marks were therefore available for the equation overall. This rubric was also collapsed



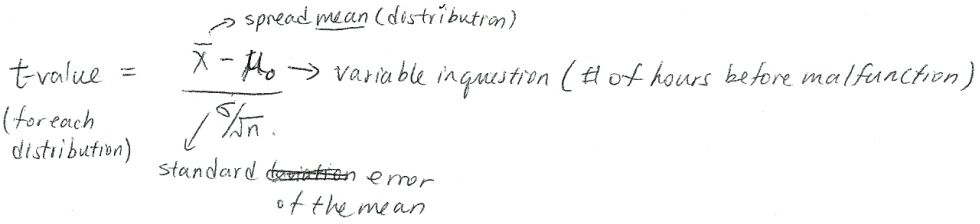
(a) Student Data 1



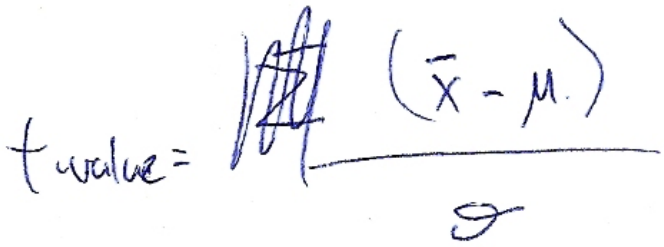
(b) Student Data 2

Figure 7.1: Sample solutions for the Recreate Data assessment by two students.

to provide a score out of 3 that measured the number of features noticed. That is, obtaining a score of 0 or 1 for a feature in the original rubric (the feature was absent or the relevant symbol was present) would obtain a 0 in the collapsed method, whereas obtain a 2 or 3 for a feature in the original rubric (the feature was present and operating in the right direction) would relate to a score of 1 in the collapsed method. The reproduced equation in Figure 7.2a demonstrates full recollection of the formula, thus obtaining a total score of 9 and a feature score of 3. The formula in Figure 7.2b received a score of 3 for each of the correct representations of central tendency and spread, but 0 for sample size. This would result in a total score of 6 and a feature score of 2.



(a) Student Formula 1



(b) Student Formula 2

Figure 7.2: Sample solutions for the Recreate Formula assessments.

7.3 Results

In all cases, ANCOVAs were carried out to determine the significance of observed differences in the data across condition (FG versus UI), with background knowledge (as measured using the CDPA) and lab section as covariates. Overall scores are summarized in Table 7.1, with the results by feature summarized in Table 7.2.

		Unguided Invention	Faded Guidance
Recreate Data	Features Score (/4)	2.86 (0.13)	2.70 (0.15)
	Pairwise Score (/3)	1.19 (0.13)	1.33 (0.14)
Recreate Formula	Equation Score (/9)	3.21 (0.36)	3.08 (0.39)
	Equation Features (/3)	0.84 (0.16)	0.82 (0.15)

Table 7.1: Average scores on the recreate data assessment by analysis type. No significant differences between groups were observed.

7.3.1 Domain Features from the Graphs

Students were able to reproduce 2.78 ± 0.10 out of 4 possible features. The UI group demonstrated significantly higher ability to recall spread in their data: UI = 0.71; FG = 0.55; Z = -2.0; p = 0.04.

7.3.2 Pairwise Comparisons in Graphs

Overall, students across conditions reproduced 1.26 ± 0.10 out of a possible 3 pairwise comparisons. A higher fraction of FG students included a pairwise contrast of sample size when recreating the graphs than the UI group: UI = 0.29, FG = 0.53; Z = 2.6; p = 0.009.

Feature	Measure	Unguided Invention	Faded Guidance
Central Tendency	Features	0.90	0.87
	Pairwise	0.52	0.55
	Equation	0.29	0.23
Spread	Features	0.71	0.55*
	Pairwise	0.38	0.25
	Equation	0.33	0.32
Sample Size	Features	0.48	0.62
	Pairwise	0.29	0.53**
	Equation	0.22	0.27
$\dagger - p < .1; * - p < .05; * * - p < .01; * * * - p < .001$			

Table 7.2: Average scores on the recreate data and equation assessment by feature and analysis type. The UI group were found to recreate spread in their data more often than the FG group (as a feature and explicitly as a pairwise comparison). The FG group were found to recreate sample size in their data more often than the UI group (as a feature and explicitly as a pairwise comparison).

7.3.3 Equation Features

On average, students scored 3.14 ± 0.27 out of 9 possible points for their equations, and reproduced 0.83 ± 0.11 features in the correct proportionalities. There was no difference between groups on performance on these items.

7.4 Conclusions

One week after completing the machine malfunction invention task described in Chapter 6, students were asked to recreate the data provided during the invention and the formula for the t-value described during the consolidation phase. The recreated data was analyzed twice: first for the presence of the domain-features; second for explicit pairwise contrasts between features.

On the recreate data task, 24% more students in the FG condition reproduced sample size as an explicit pairwise contrast than the UI group ($p < 0.01$). In apparent contradiction, 16% more students in the UI condition were able to recall the feature of spread in their recreated data than the FG group. They did not, however, reproduce this feature as a pairwise contrast any better than the FG group. This suggests that, while the UI group were better able to reproduce one of the features in the data, the FG group better understood the importance of invention strategies, especially contrasting cases and making pairwise comparisons. However, the fact that the FG group only reproduced one of the three features as pairwise comparisons more often than the UI group restricts this conclusion from being made.

However, these two results additionally contradict the results from the machine malfunction invention activity assessment, where both groups performed equally well on measures of invention quality. One explanation for this is that student performance on the invention activity did not align with the recreate data task due to the lack of scaffolding for either group during the invention activity itself. That is, the scaffolding is necessary for students to engage in successful invention strategies. The post-invention consolidation phase was, perhaps, sufficient at engaging students with the contrasts in the data and highlighting relevant features that may have been missed during the invention phase. The results of the recreate data task are consistent with the two groups extracting different lessons from the consolidation phase. In particular, UI students extracted the importance of spread, while the FG students noticed the importance of sample size.

It is interesting that the fractions of students including spread and sample size in their recreated data are significantly higher than they were in their original inventions (from 20% to approximately 50%). In contrast, their recreated equations are on approximately the same scale for these features (20 - 30 % in both cases). This suggests that the consolidation phase provided the relevant content, to each of the groups, to be able to reproduce the data the following week. It is unclear, however, what differences between the conditions caused one group to focus on spread and the other to focus on sample size.

The equations produced were analyzed similarly to the student inventions in Chapter 6, looking for the presence and correctness of domain features in the reproduced formulas. No differences were observed between groups and, on average, student formulas included less than one feature in the correct proportionality. This suggests that the invention process was insufficient in this case at connecting students to the technical qualities of the formula. This may have occurred because the consolidation phase of the machine malfunction activity included only a small practice activity, so students lacked sufficient interaction with the accepted formula.

Chapter 8

Discussion

An interactive learning environment, called the Invention Support Environment (ISE), was successfully created to deliver invention activities on the computer. The system can be scaffolded to support key invention strategies such as self-explanation, peer interactions, and iterative and cyclic inventing. It includes an equation editor component that allows students to develop mathematical equations for their inventions. The software used to build the ISE allows for easy adaptation to new invention activities. It also logs data for research purposes, such as tracking student performance on final inventions, and patterns of inventing behaviour.

The ISE was used in this thesis to examine the effects of fading domain-general metacognitive scaffolding across invention activities (FG condition) compared to unguided inventions (UI). While students under difference conditions in the study did not demonstrate significant differences in performance on technical abilities, there were some effects on measures that require deep understanding of domain features.

8.1 Discussion of Results

There was no effect for condition on the conceptual and procedural items of the statistics diagnostic. While invention activities assist with domain-level learning, the domain-general scaffolding primarily affects understanding of the functional components of the domain. This is supported by the higher performance of the FG students on the debugging items. In particular, this confirms that the faded scaffolding better supports students for understanding technical features of the canonical solutions. While both groups of students performed equally well on conceptual and procedural items, the scaffolding provided students with the tools to connect conceptual understanding to technical components of the formulas. While these results could be linked to improved development of reasoning or invention skills, these were not observed on other assessment items.

The Machine Malfunction invention activity analysis, and that of the recreated data, produced conflicting results that restrict conclusions to be drawn regarding skills acquired. In particular, the Machine Malfunction assessment was impeded due to having students work individually, rather than in pairs. The removal of peer interaction in this case added an additional variable to the analysis, so effects due to scaffolding condition could not be drawn. Improved ability to reproduce features of the domain as explicit pairwise comparisons in the recreate data assessment could suggest that the FG group had acquired a stronger understanding of the key invention strategies than the UI group. Perhaps the students in the FG group used acquired invention skills to successfully analyze the data, but lack of peer interaction and the difficulty of the task hindered their ability to form mathematical representations of the features and implement them in their inventions. However, it is unclear why they would better reproduce one feature and not the others as pairwise contrasts,

suggesting this single result is simply a statistical abnormality.

8.2 Limitations of the Experimental Design

Some of the assessments, such as the statistics diagnostic and debugging scores, revealed levels of performance and learning gains that were much lower than expected, which raises questions about the validity of the assessments themselves; namely, did the assessments measure what was intended to be measured? The least-squares fitting debugging item (Equation 4.1), for example, was perhaps technically too subtle, and students may not have noticed the misplacement of the exponent since the parentheses were all the same format (instead of embedding square or curly braces inside parentheses). Therefore, this item may have measured technical proficiency, rather than conceptual understanding. The low scores may also be representative of the significant time-delay between invention and assessment, suggesting that students struggle to transfer across time.

The amount of test validation carried out seems to have been insufficient for ensuring the quality of assessment expected. Future studies ought to conduct think-aloud interviews with students to observe their interpretations and understanding of the questions. Comparing performance on the items between more expert-like students and faculty could also provide insight into the level of knowledge and reasoning required for various levels of performance on the statistics diagnostic. That is, to assess whether an "expert" would succeed at the assessment and to what performance that success is correlated.

The original motivation to have students work independently on the near-transfer invention task was to measure individual ability and remove reliance of weaker students on stronger ones. This decision added an additional variable to the assessment and the effects of inventing individually without peer interaction may have confounded the observations. Future studies of scaffolding ought to either include this item as an observable variable with a control condition to measure its effects, or maintain consistency by including peer interaction in all inventions.

8.3 Further Research

There are various components and future changes that could be made to improve the usability and design of the ISE. For example, future versions ought to include instruction and practice activities embedded into the system, improving the transferability of the tasks to new environments and instructors. This would further ensure the control of additional variables between groups, since instructor-led discussions could vary significantly between lab sections. Development of finer-grained logging of the equation editor and inclusion of a spreadsheet component will allow more detailed analysis of students' invention behaviours, giving researchers greater insight into how students invent.

A quantitative assessment, however, of the ISE itself has yet to be carried out. In particular, it is yet to be measured whether the ISE environment improves or hinders students' abilities to invent compared to paper tasks. Analysis of student inventions using the fuel consumption task on the ISE is currently in progress.

With regards to measuring the effects of faded scaffolding during inventions, the strength of the results above provide several opportunities for further research. The results of the near-transfer invention task, where both groups received low-scaffolding during the invention, motivate comparisons of the performance of faded- and low-scaffold groups with a group that continuously receives high scaffolding. This would confirm whether scaffolding is required while inventing in order to engage students in reasoning behaviours. If the high-levels of scaffolding are necessary for successful inventing, instructors and researchers would then have to determine alternate forms of fading or support for students to acquire, encode, and transfer successful invention and reasoning strategies.

Another interesting comparison would be to analyze earlier invention tasks when the FG group received high- or medium-scaffolding compared with the UI group. This may shed light on whether higher scaffolding is always beneficial for student inventions. That is, since performance was the same on most measures of the near-transfer task (where both groups received low-scaffolding), comparing the groups during an earlier task could confirm whether the higher-level scaffolding assists students' reasoning and abilities to notice domain features [36].

In contrast, comparisons between groups during the study's final invention activity (the Lab Books activity), during which both groups received low-levels of scaffolding, may identify the effects of peer-interaction on invention ability. That is, the machine malfunction task observed students' performance on a low-structure activity, but had students working individually, rather than in pairs. This, therefore, measured the effects of removing scaffolding while also removing peer-interaction. The lab books activity also had both groups receiving low-scaffolding, but maintained the peer-interaction by having students working in pairs. Analysis of this task would, therefore, isolate the effects of scaffolding from that of peer-interaction.

In fact, analysis of all student inventions across all five main-series invention activities would discern how the faded scaffolding affected the students' performance over time, rather than just via pre- and post-study assessments. It could also reveal whether a series of UI tasks lead to improved performance over time, simply due to repeated exposure to invention activities. Unfortunately, developing coding schemes that obtain high inter-rater reliability during invention analysis is a time-consuming task. Instead, analyzing the students' implemented values for pairwise rankings for each of the invention tasks would be a relatively efficient method of examining the effects of scaffolding over time. This would allow one to observe the groups' baseline abilities and how they developed over time. If the scaffolding was indeed beneficial at the start of the study, one should be able to determine what level of scaffolding triggered the removal of differences in performance between the groups. However, the ranking data may not be sufficient without qualitative analysis of features. Since students inventions have not been found to show levels of technical proficiency [36], correct rankings could be attributed to chance. In addition, the results of the machine malfunction activity demonstrated that student rankings did not reflect whether students had noticed features of the data, as observed in the hand-coding analysis.

In order to test students' abilities to connect qualitative features to mathematical formulas a series of think-aloud interviews should be conducted where students identify features from contrasting cases and then construct a mathematical representation for each of the features. This would allow one to draw further conclusions on the debugging results. In particular, whether one group

could make such connections better than the other, or if they had better evaluating skills.

Finally, to minimize the dilution of results due to time-transfer, future studies should include the debugging questions in the ISE as transfer tasks immediately following embedded instruction and technical practice problems.

Bibliography

[1] Adobe flash professional cs5 are either registered trademarks or trademarks of adobe systems incorporated in the united states and/or other countries.

[2] Adams, W., Wieman, C., and Schwartz, D. (2008). Teaching expert thinking. Online: http://www.cwsei.ubc.ca/resources/files/Teaching_Expert_Thinking.pdf.

[3] Aleven, V., McLaren, B. M., Sewall, J., and Koedinger, K. R. (2009). A new paradigm for intelligent tutoring systems: Example-tracing tutors. *International Journal of Artificial Intelligence in Education*, 19(2):105–154.

[4] Aleven, V. A. W. M. M. and 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.

[5] Alfieri, L., Brooks, P. J., Aldrich, N. J., and Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, 103(1):1–18.

[6] Azevedo, R., Cromley, J., and Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia. *Contemporary Educacational Psychology*, 29:344–370.

[7] Azevedo, R. and Hadwin, A. (2005). Scaffolding self-regulated learning and metacognition – implications for the design of computer-based scaffolds. *Instructional Science*, 33:367–379. 10.1007/s11251-005-1272-9.

[8] Belenky, D. (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*, pages 1163–1168.

[9] Buffler, A., Allie, S., and Lubben, F. (2001). The development of first year physics students' ideas about measurement in terms of point and set paradigms. *International Journal of Science Education*, 23(11):1137–1156.

[10] Bulu, S. and Pedersen, S. (2010). Scaffolding middle school students' content knowledge and ill-structured problem solving in a problem-based hypermedia learning environment. *Educational Technology Research and Development*, 58:507–529. 10.1007/s11423-010-9150-9.

[11] Carolan, J. (2011). Personal communication with Dr. Jim Carolan, a professor emeritus in the Department of Physics and Astronomy at the University of British Columbia. Dr. Carolan has been collecting diagnostic data on various first-year courses at UBC for the past 20 years.

[12] Chase, C. C., Shemwell, J. T., and Schwartz, D. L. (2010). Explaining across contrasting cases for deep understanding in science: an example using interactive simulations. In *Proceedings of the 9th International Conference of the Learning Sciences - Volume 1*, ICLS '10, pages 153–160. International Society of the Learning Sciences.

[13] Chase, W. G. and Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4(1):55 – 81.

[14] Chi, M. T., Leeuw, N. D., Chiu, M.-H., and Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439 – 477.

[15] Choi, I., Land, S., and Turgeon, A. (2005). Scaffolding peer-questioning strategies to facilitate metacognition during online small group discussion. *Instructional Science*, 33(5-6):483–511.

[16] Dabbagh, N. and Kitsantas, A. (2005). Using web-based pedagogical tools as scaffold for self-regulated learning. *Instructional Science*, 33(5-6):513–540.

[17] Day, J. and Bonn, D. (2011). Development of the concise data processing assessment. *Physical Review Special Topics - Physics Education Research*, 7(1).

[18] Day, J., Nakahara, H., and Bonn, D. (2010). Teaching standard deviation by building from student invention. *The Physics Teacher*, 48(8):546–548.

[19] De Jong, T. and Van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, 68(2):179–201.

[20] Elby, A. (1999). Another reason that physics students learn by rote.

[21] Gray, K. E., Adams, W. K., Wieman, C. E., and Perkins, K. K. (2008). Students know what physicists believe, but they don't agree: A study using the class survey. *Physical Review Special Topics - Physics Education Research*, 4(2).

[22] Hadwin, A., Wozney, L., and Pantin, O. (2005). Scaffolding the appropriation of self-regulatory activity; a socio-cultural analysis of changes in teacher-students discourse about a graduate research portfolio. *Instructional Science*, 33(5-6):413–450.

[23] Hestenes, D., Wells, M., and Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30(3):141–158.

[24] Kapur, M. (2008). Productive failure. *Cognition and Instruction*, 26(3):379–424.

[25] Kapur, M. (2010). A further study of productive failure in mathematical problem solving: unpacking the design components. *Instructional Science*, 39:561–579. 10.1007/s11251-010-9144-3.

[26] Kapur, M. and Kinzer, C. (2009). Productive failure in cscl groups. *International Journal of Computer-Supported Collaborative Learning*, 4:21–46. 10.1007/s11412-008-9059-z.

[27] Kirschner, P. A., Sweller, J., and 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.

[28] Koedinger, K., Anderson, J., Hadley, W., and Mark, M. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8(1):30–43.

[29] Koedinger, K. R., Aleven, V., Heffernan, N., McLaren, B., and Hockenberry, M. (2004). Opening the door to non-programmers: Authoring intelligent tutor behavior by demonstration. In Lester, J. C., Vicari, R. M., and Paraguaçu, F., editors, *Intelligent Tutoring Systems*, volume 3220 of *Lecture Notes in Computer Science*, pages 7–10. Springer Berlin / Heidelberg. doi:10.1007/978-3-540-30139-4_16.

[30] Li, S. Personal communication with Salena Li, undergraduate program coordinator in the Department of Physics and Astronomy at the University of British Columbia.

[31] Mathan, S. A. and Koedinger, K. R. (2005). Fostering the intelligent novice: Learning from errors with metacognitive tutoring. *Educational Psychologist*, 40(4):257–265.

[32] Puntambekar, S. and Stylianou, A. (2005). Designing navigation support in hypertext systems based on navigation patterns. *Instructional Science*, 33(5-6):451–481.

[33] Roll, I., Aleven, V., and Koedinger, K. (2009). Helping students know 'further' - increasing the flexibility of students' knowledge using symbolic invention tasks. In *The 31st Annual Conference of the Cognitive Science Society, Cognitive Science Society.*, volume 1, pages 1169–1174, Austin, TX. Cognitive Science Society.

[34] Roll, I., Aleven, V., and Koedinger, K. (2010). The invention lab: Using a hybrid of model tracing and constraint-based modeling to offer intelligent support in inquiry environments. In Aleven, V., Kay, J., and Mostow, J., editors, *Intelligent Tutoring Systems*, volume 6094 of *Lecture Notes in Computer Science*, pages 115–124. Springer Berlin / Heidelberg. 10.1007/978-3-642-13388-6_16.

[35] Roll, I., Aleven, V., and Koedinger, K. R. (To Appear). Outcomes and mechanisms of transfer in invention activities. In *Cognitive Science*.

[36] Roll, I., Holmes, N., Day, J., and Bonn, D. (in-press). On guided invention activities that support scientific reasoning and domain learning. *Instructional Science*.

[37] Schoenfeld, A. H. (1987). What's all the fuss about metacognition. In Schoenfeld, A. H., editor, *Cognitive science and mathematics education*, pages 189–216. Lawrence Erlbaum Associates, Hillsdale.

[38] Schwartz, D. L. and Bransford, J. D. (1998). A time for telling. *Cognition and Instruction*, 16(4):475–522.

[39] Schwartz, D. L. and 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.

[40] Strand-Cary, M. and Klahr, D. (2008). Developing elementary science skills: Instructional effectiveness and path independence. *Cognitive Development*, 23(4):488 – 511. Scientific reasoning – Where are we now?

[41] Sweller, J. (2009). What human cognitive architecture tells us about constructivism. In Tobias, S. and Duffy, T. M., editors, *Constructivist Instruction: Success or failure?*, pages 127–143. Routledge/Taylor & Francis Group, New York, NY, US.

[42] Taylor, J. L., Smith, K. M., van Stolk, A. P., and Spiegelman, G. B. (2010). Using invention to change how students tackle problems. *CBE Life Sci Educ*, 9(4):504–512.

[43] VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16(3):227–265.

[44] White, B. Y. and Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction*, 16(1):3–118.

Appendix A

Statistics Diagnostic

A.1 Pre- and Post-Test Items

1. A battery company tested a new rechargeable battery, Battery X, and plotted lifetime vs. charging-time. They created a statistical tool to calculate the best-fit linear relation that matches their data, as shown in Figure A.1.

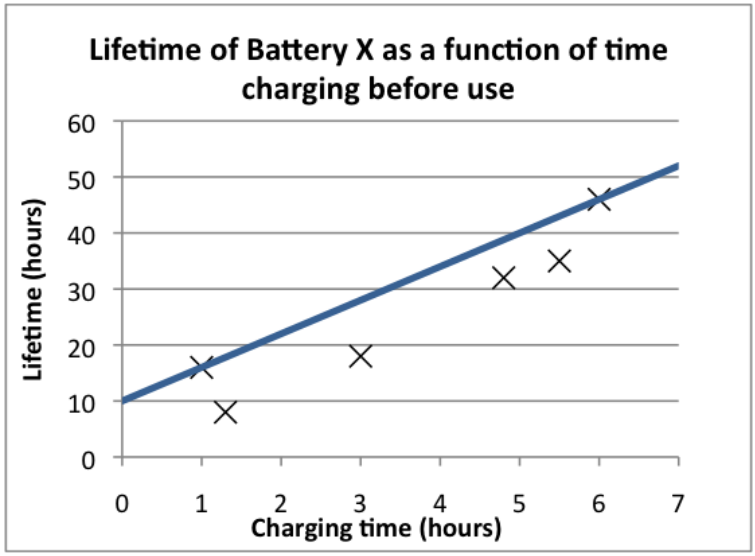


Figure A.1: Battery data for Question 1.

Do you think the straight line represents the best straight-line fit for the data? Please choose the best answer.

(a) Yes

(b) No, the line of best fit lines should go through as many points as possible

(c) No, the line of best fit should balance the number of points above and below the line

(d) No, the line of best fit should minimize the distance between the line and the points

(e) None of the above

2. Bob measures the length of a stick to be (A $\pm$ 2) mm. Alice measures the length of the same stick to be (B $\pm$ 1) mm. Which statement is most reasonable?

(a) The stick is closer to A mm than it is to B mm

(b) The stick is (A + B)/2 mm

(c) The stick is closer to B mm than it is to A mm

(d) It is not possible to determine this without knowing the values of A and B

The following information, as well as Table A.1 and Figure A.2 apply to questions 3 and 4. For an experiment examining Ohms Law (V= I R), students made measurements of current, I, in a series circuit with constant resistance as they varied the voltage from 0 to 10 V. They used a best fitting line (y = 1.889x) to determine the resistance used in the circuit.

Voltage (V)	Current (mA)
0.1	0.2
0.5	0.5
0.7	2.4
1.4	3.1
1.9	2.8
2.3	5.4
2.8	4.9
3.3	7.3
3.8	7.5
4.2	8.9
4.9	9.7
5.4	9.3
5.9	11.4
6.5	11.2
7.0	12.0
7.3	14.6
8.1	15.4
8.7	17.0
9.3	16.5
9.9	19.3

Table A.1: Measurements of voltage and current through a circuit, for questions 3 and 4.

3. What is the least-squares residual error of the best fit line to the data in Figure A.2, given the corresponding data in Table A.1?

   (a) 0.46 mA^2

   (b) 0.59 mA^2

   (c) 11.85 mA^2

   (d) 13.62 mA^2

   (e) None of the above

4. What is the uncertainty in the value of the slope?

   (a) 0.031 mA/V

   (b) 0.0018 mA/V

   (c) 0.45 mA/V

   (d) 0.77 mA/V

   (e) None of the above

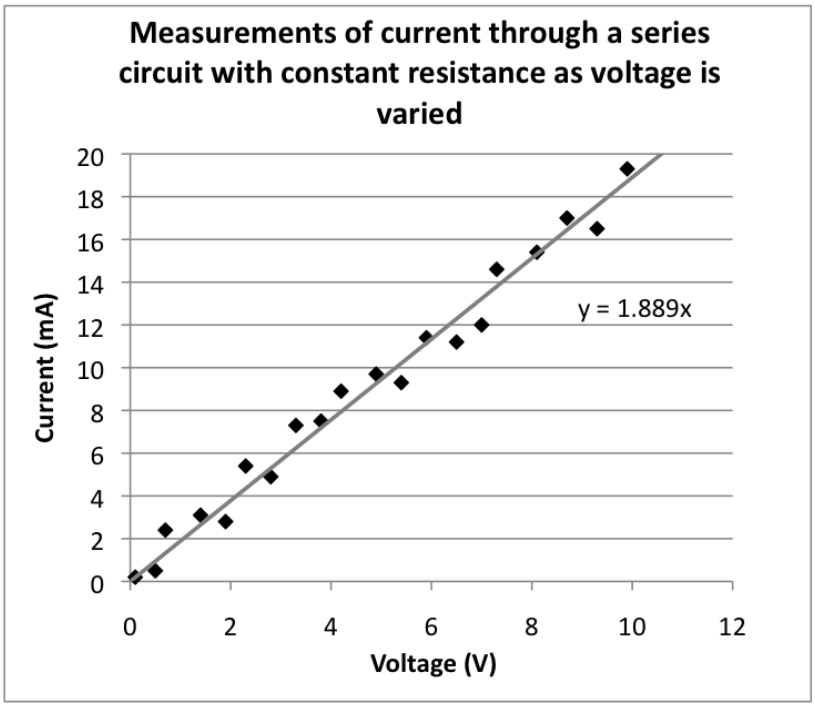


Figure A.2: Measurements and best-fitting linear relationship of current as voltage is changed in a series circuit with constant resistance.

5. Two students, working together on a lab experiment, were asked to measure the period of the same pendulum. They decided to each use a different method to do so. The first student reported the period to be (3.3 ± 0.2) seconds and the second reported (2.8 ± 0.6) seconds, each using several trials to get their values. What would you record as the most likely value if you were to repeat the experiment many more times?

   (a) 2.98 s

   (b) 3.05 s

   (c) 3.18 s

   (d) 3.25 s

   (e) None of the above

A.2 Debugging Items (Post-Only)

The following three questions (6, 7 and 8) are of the same form. Below is an example of how one may answer such a question.

Question: Is the following formula a valid way to capture the variability of a data set?

$$x_{max} - x_{min}$$

Answer: No, since it looks at the range but ignores all other numbers.

6. Is the following formula valid for evaluating the quality of a best-fit line? Justify your answer.

$$\chi^2 = \frac{1}{N}\sum_{i=1}^{N}\Big(y_i - f(x_i)\Big)^2 \tag{A.1}$$

7. Is the following formula a reasonable way to calculate a weighted mean of values x_i with associated uncertainties δx_i? Justify your answer.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i \delta x_i}{\sum_{i=1}^{N} \delta x_i} \tag{A.2}$$

8. Is the following formula a valid way to calculate the uncertainty, ?m, in the slope, m, of a best-fit line with zero intercept? Justify your answer.

$$\delta m = \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{y_i}{x_i} - m\Big)^2 \tag{A.3}$$

Appendix B

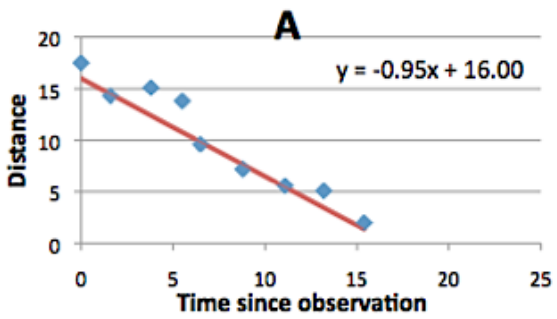
Invention Activities Used in the Study

The following are brief descriptions of the individual invention activities used in the study. In particular, the cover story, graphs and data for each task are provided.

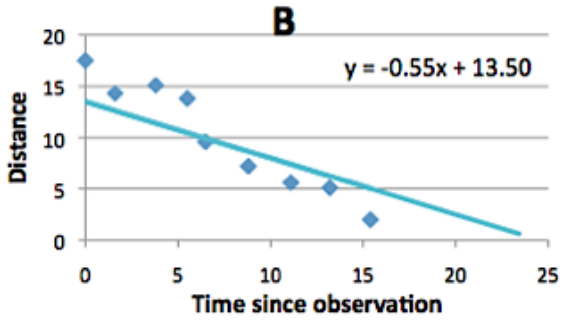
B.1 Planet Phaedra

The goal of this activity is for students to explore what makes a line of best fit and how fitting lines can be evaluated. After the activity, students receive instruction about linear least squares fitting, and the chi-square test (Equation B.1).

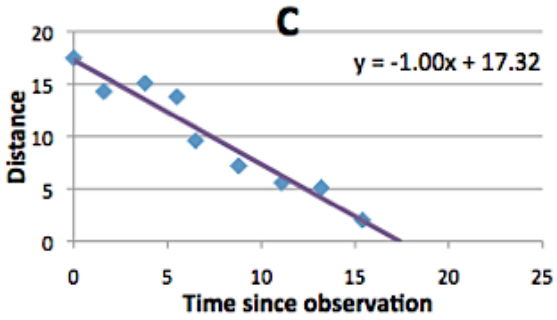
$$\chi^2 = \frac{1}{N}\sum_{i=1}^{N}\Big(y_i - f(x_i)\Big)^2 \tag{B.1}$$



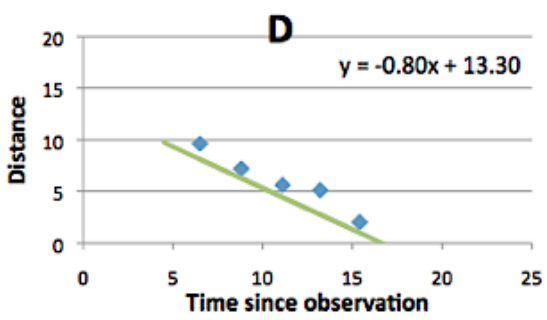
(a) Advisor A's Best Fit model



(b) Advisor B's Best Fit model



(c) Advisor C's Best Fit model



(d) Advisor D's Best Fit model

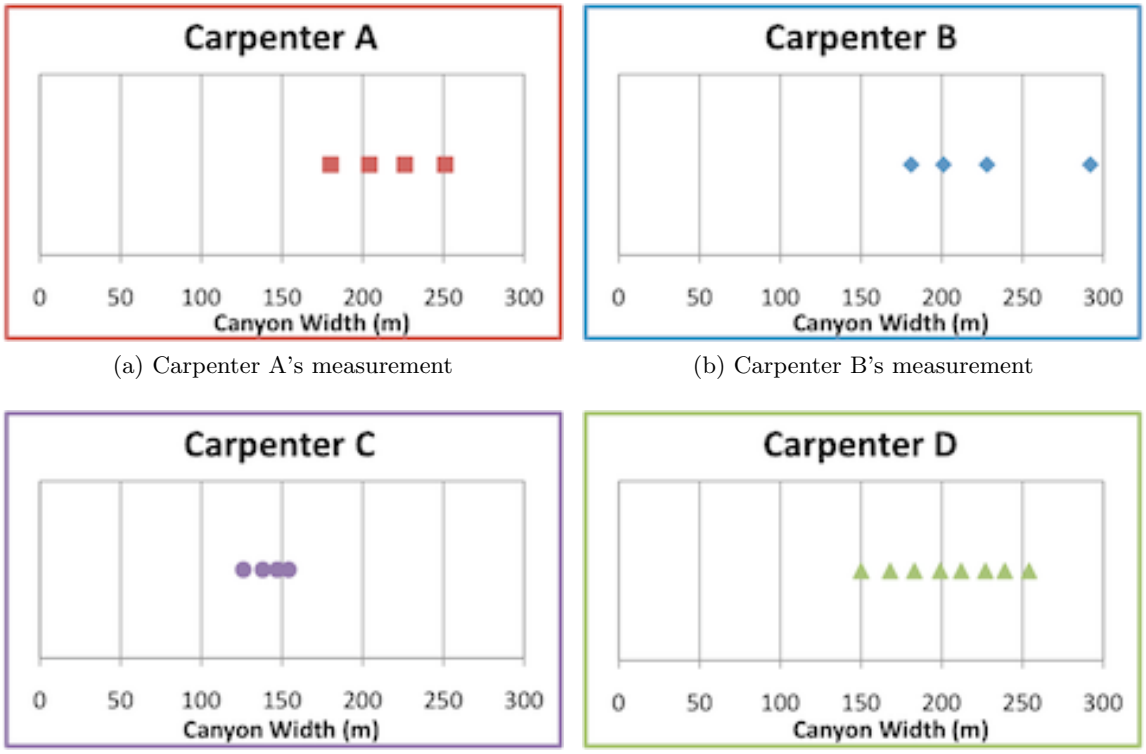
Figure B.1: The data for the Planet Phaedra invention activity. The four graphs represent the best fit lines provided by four advisors attempting to predict when an asteroid will collide with the planet Phaedra

The task describes a set of measurements of an asteroid's position over time as it approaches the planet Phaedra. Four advisors have fit models to the data in order to determine when the asteroid would collide with the planet. The students are required to invent a formula to quantitatively determine which of the advisors' fit models best represents the data provided. The four data sets are provided in Figures B.1.

B.2 The Not-so-Grand Canyon

The goal of this activity is for students to explore measurement uncertainties and how they should be taken into account. After the activity, students receive instruction about weighted averages (Equation B.2).

$$\overline{x_w} = \frac{\sum_{i=1}^{N} \frac{x_i}{\delta x_i^2}}{\sum_{i=1}^{N} \frac{1}{\delta x_i^2}} \tag{B.2}$$



(a) Carpenter A's measurement

(b) Carpenter B's measurement

(c) Carpenter C's measurement

(d) Carpenter C's measurement

Figure B.2: The data for the Not-so-Grand Canyon invention activity. The four graphs represent measurements of the width of a canyon by four carpenters. The goal of the task is to use their data to determine the true width of the canyon.

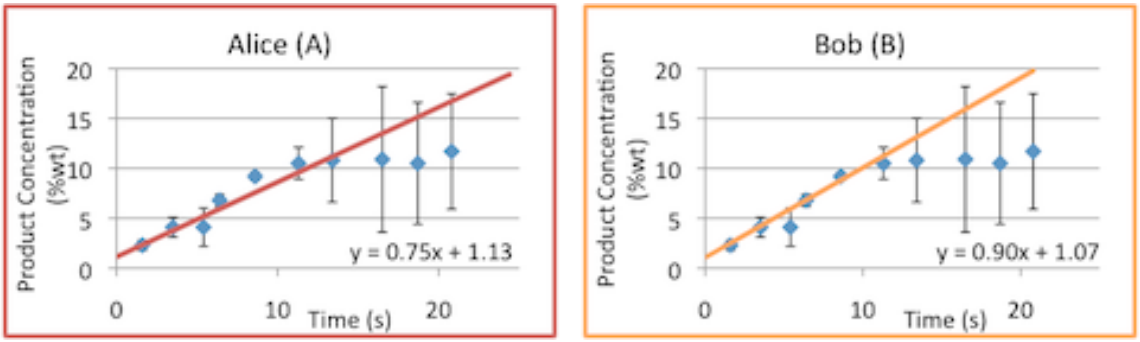
The task describes measurements of the width of a canyon by four village carpenters. It is accepted that some measurements are better than others. The students are required to invent a formula to determine the true width of the canyon using the carpenter's measurements. The four

data sets are provided in Figures B.2.

B.3 Glucose Oxidation

The goal of this activity is for students to explore measurement uncertainties and how they affect fitting to data. After the activity, students receive instruction about weighted linear least squares (Equation B.3).

$$\chi^2 = \frac{1}{N} \frac{\sum_{i=1}^{N} \left(\frac{y_i}{\delta y_i^2} - \frac{f(x_i)}{\delta y_i^2} \right)^2}{\sum_{i=1}^{N} \frac{1}{\delta y_i^2}} \tag{B.3}$$



(a) Alice's fit

(b) Bob's fit

(c) Carlita's fit

(d) Damien's fit

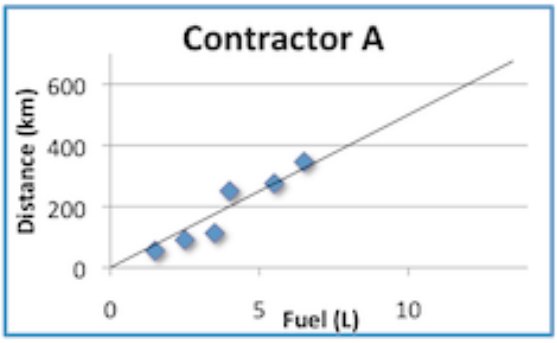
Figure B.3: The data for the Glucose Oxidation invention activity. The four graphs represent different fits to the same data. The goal of the task is to use their data to quantitatively determine which of the fits best represents the data.

The task describes measurements of the concentration of products from a glucose oxidation reaction in chemistry lab experiment. Alice, Bob, Carlita and Damien are working together on the lab and have each come up with a different fit to the data. Students are required to invent a formula to quantitatively determine which of the fits best represent the data. The four data sets are provided in Figures B.3.

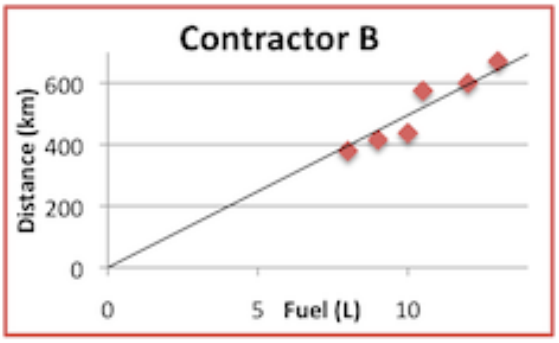
B.4 Fuel Consumption

The goal of this activity is for students to explore what properties of data improve uncertainty in the slope of a best-fit line. After the activity, students receive instruction about the slope uncertainty for a best-fit line whose intercept is locked at zero (Equation B.4).

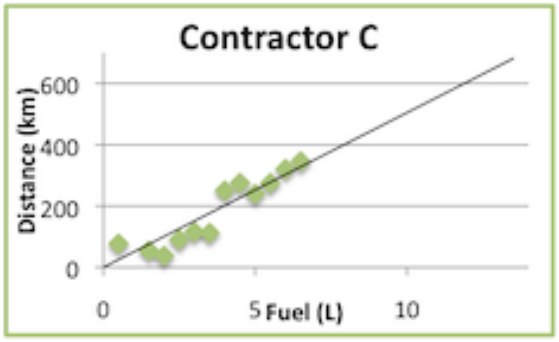
$$\delta m^2 = \frac{1}{N} \frac{\sum_{i=1}^{N} \Big(y_i - f(x_i)\Big)^2}{\sum_{i=1}^{N} x_i^2} \tag{B.4}$$



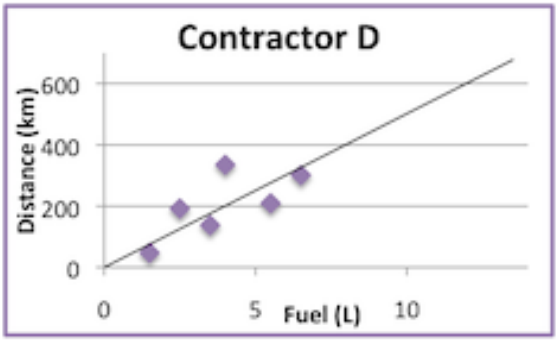
(a) Contractor A's measurements



(b) Contractor B's measurements



(c) Contractor C's measurements



(d) Contractor D's measurements

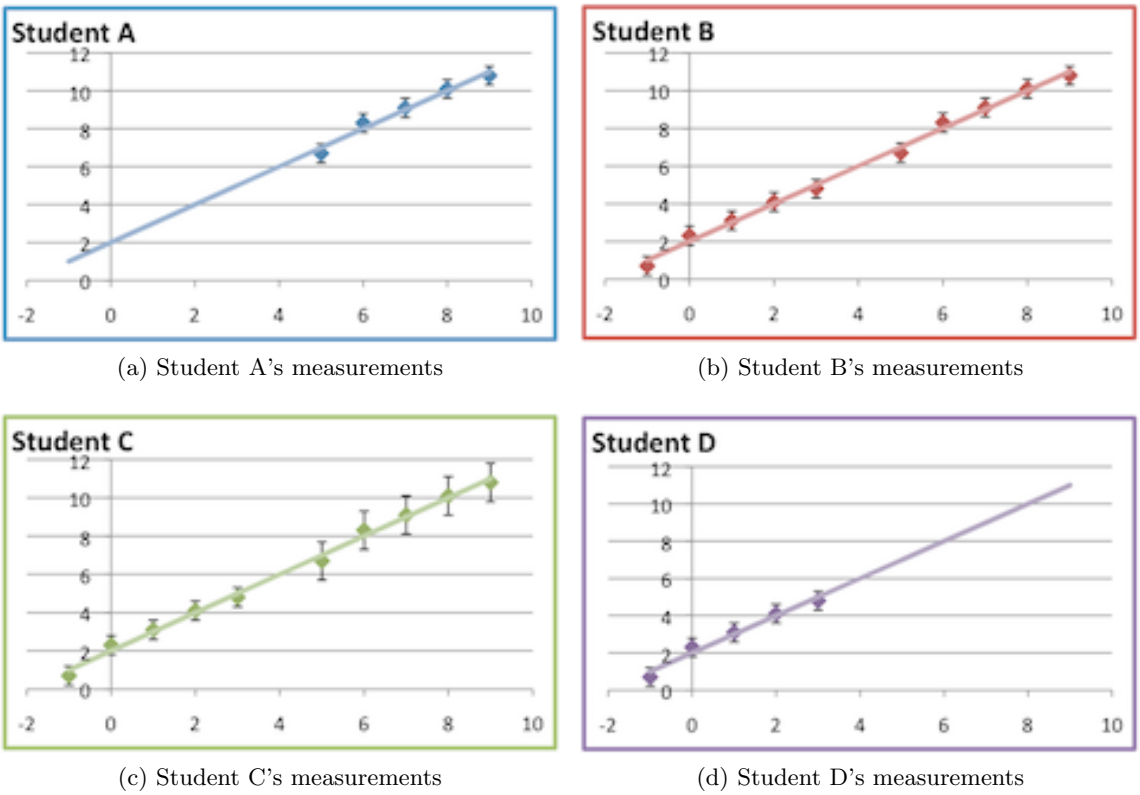
Figure B.4: The data for the Fuel Consumption invention activity. The four graphs represent different measurements to determine the fuel efficiency of a vehicle. The goal of the task is to quantitatively determine which of the measurements does a better job at measuring the slope of the best-fit line.

The task describes four contractors who have made measurements on the fuel efficiency of a vehicle. Although each contractor has measured an efficiency of 50 km/L, representing the slope of the line, their methods of measurement differ significantly. Students are required to invent a formula to quantitatively determine which of the sets of measurement is most accurate by inventing a formula for the uncertainty in the slope of the best-fit lines. The four data sets are provided in Figures B.4.

B.5 Lab Books

The goal of this activity is for students to explore measurement uncertainties and how they affect the slope of a best-fit line. After the activity, students receive instruction about the slope uncertainty for a best-fit line with a non-zero intercept and uncertainties in the data points (Equation B.5).

$$\delta m^2 = \frac{1}{N} \frac{\sum_{i=1}^{N} \left(\frac{y_i}{\delta y_i^2} - \frac{f(x_i)}{\delta y_i^2} \right)^2}{\sum_{i=1}^{N} \frac{x_i^2}{\delta y_i^2}} \tag{B.5}$$



(a) Student A's measurements

(b) Student B's measurements

(c) Student C's measurements

(d) Student D's measurements

Figure B.5: The data for the Lab Books invention activity. The four graphs represent different measurements made by four different students with the same best-fit line. The goal of the task is to quantitatively determine which of the measurements does a better job at measuring the slope of the best-fit line, taking into account the non-zero intercept and the data point uncertainties.

The task describes a TA marking four lab reports with graphs that had the same linear relationship but different data values. Students are required to invent a formula to quantitatively determine which of the sets of measurement is most accurate by inventing a formula for the uncertainty in the slope of the best-fit lines. This differed from the Fuel Consumption activity in that the intercept was not locked at zero and the data points had uncertainties. The four data sets are provided in Figures B.5.