The Difficulty Factors Approach to the Design of Lessons in Intelligent Tutor Curricula

Ryan S.J.d. Baker, *Learning Sciences Research Institute, University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham NG8 1BB, UNITED KINGDOM. ryan@ educationaldatamining.org*

Albert T. Corbett, *Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213, USA, corbett@cmu.edu*

Kenneth R. Koedinger, *Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213, USA, koedinger@cmu.edu*

Abstract. We present an approach to designing intelligent tutoring systems, termed the Difficulty Factors Approach. In this approach, the designer investigates, at each iteration of the design cycle, which skills and concepts are difficult for students, and what factors underlie those difficulties. We show that this approach complements existing design principles, producing data that helps designers apply principles in context. We also show that by continuing to investigate student difficulties throughout the design process, it is possible to discover difficulty factors initially obscured by other difficulty factors. We give an example of the application of the Difficulty Factors Approach in the context of the development of a cognitive tutor lesson on scatterplots.

Keywords. Intelligent Tutoring Systems, Instructional Design Principles, Evaluation of AIED Systems, Difficulty Factors Assessments

INTRODUCTION

Within most educational domains, the designer of an interactive learning environment has a seemingly limitless number of choices to make. Such a designer must decide what sorts of skills and concepts the system should attempt to help students learn, and must decide what sorts of learning support the system should offer.

In recent years, considerable progress has been made in learning how systems can support students. Within this journal and related publication venues, a large number of educationally successful systems have been presented, for a wide variety of domains, providing designers with a rich set of examples to draw from in developing a new system (cf. Simon, 1996). In addition, the experiences gained by developing these systems have been distilled into several sets of usable and useful design principles (e.g. Merrill & Reiser, 1993; Anderson, Corbett, Koedinger, & Pelletier, 1995; Linn, 1995; Baumgartner & Bell, 2002; Quintana et al, 2004; van Lehn et al, 2005). Such design principles have even been incorporated into authoring systems (cf. Murray, Woolf, & Marshall, 2004; Koedinger, Aleven, Heffernan, McLaren, & Hockenberry, 2004), reducing the amount of work needed to develop systems that conform to validated principles for good design.

However, having a well-designed and empirically validated set of principles – or a high-quality authoring system – is not sufficient to ensure that the resulting tutoring system will be educationally effective. This is because design principles must still be applied in a system designed for a specific

domain and population. Design principles tell us how to design educational systems that help students learn specific skills and concepts; they do not tell us which skills and concepts to focus on. Take, for example, the principle proposed by Anderson et al. (1995), that intelligent tutor designers should "Communicate the goal structure underlying the problem solving" and make explicit the steps which are implicit in the problem-solving process. This principle, unfortunately, does not tell us *which* implicit steps to make explicit. A process can be broken down into steps of a finer grain-size almost endlessly. Hence, in order to follow this principle, the tutor designer first needs to decide which parts of the goal structure to communicate and make explicit. One reasonable heuristic is to attempt to communicate the parts of the goal structure that present difficulty for students. However, to do this, the designer must know which parts of the goal structure present difficulty.

Before we discuss how the designer can determine which parts of the goal structure present difficulty, it is worth pointing out that this issue does not just apply to a single principle. Consider, for example, Quintana et al's (2004) principle, "Use descriptions of complex concepts that build on learners' intuitive ideas". The designer needs to know which complex concepts to describe before he or she can implement this principle, and one way to implement this principle is by describing the complex concepts which provide particular difficulty for students.

In general, in order to properly fulfill a design principle that recommends learning support, the designer needs to know where learning support will be needed. Hence, our developer will need to ask: What causes difficulty for students in the domain of interest?

There are a number of ways to obtain an answer to this question. The quickest method is for the designer to use his/her intuition. However, although intuition is an important part of any design process, it is not a good way to determine what will be difficult for students. Educational designers are frequently too different from the students who will use their system to know what the student will find difficult (cf. Koedinger, 2002). Even experienced teachers are often unable to guess what types of problems or aspects of a problem-solving process will present difficulty for their students (Nathan & Koedinger, 2000).

Another approach to determining which steps will be difficult is rational task analysis (cf. Reigeluth, Bunderson, & Merrill, 1994). However, rational task analysis shares the same risk as intuition. If students use problem-solving strategies which are different from those modeled in task analysis, the designer may provide support for the right parts of the wrong problem-solving strategies.

Given these concerns, a growing number of educational researchers have sought to determine what parts of the problem-solving process are most challenging for students through empirical methods for task analysis (cf. Lovett, 1998). By explicitly gathering data on students' processes and difficulties, designers have a better chance of understanding what factors will cause difficulty for students and designing systems that offer support to students where it is most needed.

At this point, a considerable amount is known about which methods are effective for eliciting understanding about student thinking (cf. Ericsson & Simon, 1993, Heffernan & Koedinger, 1998). However, it is still an open question how these methods can be most appropriately integrated into the overall process of designing interactive learning environments.

For instance, one common view is that empirical task analysis should be used prior to designing a learning environment, in order to inform the later design. While this view makes a clear place for empirically investigating student difficulties, it does not take into account the possibility that analyses conducted in advance may miss difficulty factors which cannot be expressed until another difficulty factor has been surmounted with the tutoring system's help -- for example, if a student does not know how to start a car, we are unlikely to discover that the student does not know which pedal is the brake.

An alternate approach is to make the study of student difficulty factors a focus of the entire process of designing a learning environment, attempting at each stage of design to see whether the known difficulty factors are being addressed, and whether new, previously hidden difficulty factors are now salient. We term this approach the Difficulty Factors Approach to educational design.

A focus on difficulty factors has three types of antecedents in the educational design literature. The first is in the use of a specific type of experiment, termed a Difficulty Factors Assessment (DFA) (cf. Heffernan & Koedinger, 1998; Koedinger & Nathan, 2004), to determine student difficulty factors prior to design. DFAs are an essential part of a difficulty factors approach to design; however, our approach extends the focus on difficulty factors considerably beyond just the use of this type of experiment. We will discuss DFA studies in detail in the next section.

A second type of antecedent is in work towards modeling and remediating misconceptions and bugs, incorrect concepts or skills which make it more difficult for students to learn more correct concepts or skills (e.g. Clement, 1982, vanLehn, 1990). Existing methods for studying misconceptions and bugs integrate easily into an approach to design which centers on which factors present the greatest difficulty for students.

A third antecedent to the difficulty factors approach to design is in work towards developing frameworks to help designers choose which types of learning support or design principles to use, based on what types of cognitive challenges are present in the domain of interest (de Jong and van Joolingen, 1998; Quintana et al, 2004). As such a framework depends on understanding what students' cognitive challenges are, this approach combines naturally with a design approach that help designers develop better understanding of students' cognitive challenges.

In this paper, we will discuss the difficulty factors approach to design, focusing our discussion on the design of an individual lesson in a cognitive tutor curriculum. The approach we articulate is applicable much more broadly – we believe that it can be applied to any learning environment which attempts to help students learn a specific curriculum of material. However, we have thus far only applied this approach to the design of cognitive tutor lessons, and will restrict our discussion to this context.

In order to comprehensively illustrate the approach, we will focus our discussion on a specific program of research which produced a cognitive tutor lesson (Koedinger, Anderson, Hadley, & Mark, 1997) which helped students learn to create scatterplots of data. This is not the only context which we have used a difficulty factors approach to design – in particular, DFAs and a general focus on difficulty factors have been key to our research group's work to design many tutor lessons, including lessons on algebraic symbolization (Heffernan & Koedinger, 1998; Heffernan, 2001), geometric proof (Aleven et al, 1998; Aleven & Koedinger, 2002), and story problem solving (Koedinger & Nathan, 2004; Koedinger & Corbett, 2006). However, by discussing our approach in the context of the development of a single tutor lesson, we hope to be able to concretely show how the approach is conducted, how difficulty factors are researched at each stage of the design process, and how focusing on difficulty factors leads to a measurably effective tutor lesson.

During the course of developing this lesson, difficulty factors research helped us to discover a characteristic misconception, present across many populations, that led many students to create scatterplots which incorporated features from another type of data representation, resulting in a representation that could not be used for important types of data interpretation. To the best of our knowledge, this misconception had not been addressed in prior curricula where students create scatterplots. Studying this misconception enabled us to create a scaffold which helped students learn to distinguish the informational properties of scatterplots from those of other representations. Continuing to collect data on student difficulty factors even as we evaluated our first prototype led to the discovery of additional student difficulty factors, which were addressed in later stages of our design process. The final result of this process was a cognitive tutor lesson that led to high learning gains, across a diversity of educational settings.

The Difficulty Factors Approach, and Difficulty Factors Assessments (DFA)

Many techniques can be used to figure out which skills or types of problems are most difficult for students. One technique that our research group has found particularly useful is the Difficulty Factors Assessment (DFA) (Heffernan and Koedinger, 1998). A DFA focuses on systematically varying problem characteristics in order to figure out which skills are most difficult for students. A DFA is not focused on enumerating or understanding all of the errors any student makes (e.g. Cohen, Smith, Chechile, Burns, & Tsai, 1996; vanLehn, 1990); instead the goal of a DFA is to determine which factors present the greatest difficulties for students.

We classify three types of studies as DFAs. The first type of DFA, called a "Sub-Domain Selection DFA", compares between problems which differ in the exact set of skills exercised but which do not differ in other ways, in order to determine which set of skills presents the greatest overall challenge to student (e.g. Siegler, 1976; Heffernan & Koedinger, 1998; e.g. Aleven et al, 1998; Koedinger & Nathan, 2004). Sub-Domain Selection DFAs are useful when a domain is not yet well known, and it is not clear what parts of the domain should be focused on in instructional design. In addition, analyzing the specific pattern of errors on the domain sub-sections found to be difficult by a Sub-Domain Selection DFA can help focus the design of subsequent DFAs.

The second type of DFA, called a "Strategy Selection DFA", compares between different strategies within a specific representative problem, either by requesting that students use specific problem-solving strategies (e.g. Rittle-Johnson & Koedinger, 2001) or by observing which strategies students naturally choose and identifying which strategies are associated with substantially better or worse performance (e.g. Siegler, 1987; Koedinger & Nathan, 2004). When it is already known that a specific problem is difficult (either from prior research or from the results of a Sub-Domain Selection DFA), a Strategy Selection DFA can help determine what approaches are effective for solving this problem, either so that those strategies can be taught or used as the basis of bridging strategies, which build upon intuitive strategies in order to help students learn more sophisticated strategies later (cf. Koedinger, 2002)

The third type of DFA, called a "Step Support DFA", investigates whether offering support on specific problem-steps (which are thought during study design to be especially difficult for students) makes the entire problem-solving process more accessible to students, within a specific representative problem and solution strategy (e.g. Vygotsky, 1978; Rittle-Johnson & Koedinger, 2005). When it is already known that a specific problem is difficult (either from prior research or from the results of a Sub-Domain Selection DFA), and the appropriate strategy for solving that problem is known (either because it has been task-analytically determined that there is a single feasible main path to solving the problem, or because the most appropriate strategy has been determined using a Strategy Selection DFA), a Step Support DFA can help determine what parts of the selected strategy are most difficult for students.

We present two studies in this paper which we label as DFAs. Study DFA1 is a Sub-Domain Selection DFA which was used to select the sub-domain focused on in the design program discussed here; the data generated in this DFA is then used in an error analysis. Due to the nature of the domain selected (scatterplot generation) and the difficulty factors discovered in study DFA1, a Strategy Selection DFA is not immediately needed. Instead, study DFA1 is followed directly by study DFA2, a Step Support DFA used to determine which steps of the problem-solving process need extra support.

Using a Difficulty Factors Centered Design Process to Design a Specific Cognitive Tutor Lesson

In the following sections, we will discuss our work to develop a cognitive tutor lesson on scatterplots, using a difficulty factors centered design process. Our research on that lesson took place during the development of a year-long curriculum; in this discussion, we focus on a single lesson (2-3 class periods long) in order to show how the difficulty factors approach to designing tutor lessons works in practice.

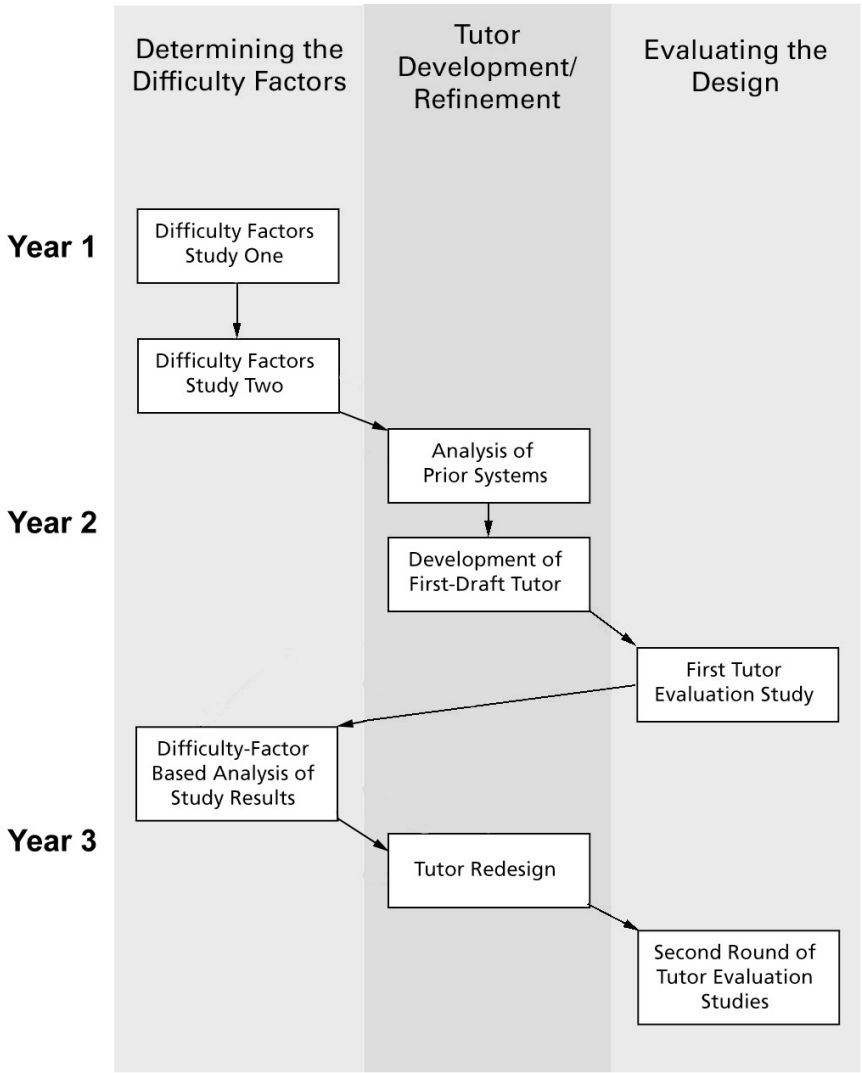


Fig. 1. The process of developing our scatterplot tutor lesson

Within the difficulty factors approach to design, the designer engages in three activities: studying difficulty factors, developing a cognitive tutor lesson to address those difficulty factors, and evaluating the tutor lesson to see how effectively it made the difficult skills and concepts accessible to students. In many cases, evaluating a tutor lesson gives new evidence as to what the most important difficulty factors are; there is growing consensus in both educational design (Collins et al, 2004) and other areas of design (McConnell, 1996; Kelley, 2001) that creating and evaluating artifacts often produces new knowledge that suggests further investigation into user/student needs.

Our process for creating a cognitive tutor lesson on scatterplots (shown in Figure 1) was as follows: we first conducted two exploratory studies to learn, to a first level of approximation, what factors produced difficulty for students in this domain. Both studies were Difficulty Factors Assessments (DFA). Each study was conducted in a mixture of urban and suburban classrooms.

Given our first-draft understanding of the student difficulty factors, we then developed a first-draft cognitive tutor lesson, to scaffold students in learning the skills and concepts which our research suggested were difficult for students. After we completed the first deployment of the cognitive tutor lesson, we analyzed the errors students made before and after using the software, and discovered an additional difficulty factor masked by previous difficulty factors.

After analyzing the data from this lesson, we had a richer understanding of the difficulty factors in this domain. We used this understanding to develop a second-draft cognitive tutor lesson. We evaluated this refined tutor lesson in suburban schools and with homeschool students and determined that this updated version effectively addressed all of the known difficulty factors to at least some degree, leading to excellent learning.

One important aspect of our design approach was the choice of students from different settings (i.e. students in urban classrooms, suburban classrooms, and homeschools) in different phases of the study. Using students with different prior educational experiences is not a type of difficulty factors assessment in and of itself (unless the goal is to study *how* different prior educational experiences affect student difficulty factors). However, using students with different prior educational experiences makes it more likely that the model of difficulty factors developed will describe the broadest possible set of students, and that the tutor lesson developed using that model will be effective when used by students outside of the original populations studied. Conducting a set of difficulty factors assessments within a single very homogenous population creates the very real risk of designing a system that is only effective within a single school or educational setting.

In the following sections, we will go into greater detail about each of these studies, and the cognitive tutor lessons developed at each stage of the process. We will also discuss the methods we used, and how these methods fit in to a difficulty factors approach to design.

PHASE 1: INVESTIGATING STUDENTS' DIFFICULTIES IN CREATING SCATTERPLOTS OF DATA

The cognitive tutor lesson which we will discuss was developed as part of a broad, standards-based (cf. National Council of Teachers of Mathematics, 2000) cognitive-tutor based curriculum for middle school mathematics (see Koedinger, 2002 for a fuller discussion of this course). The goal of the lesson was to help students learn how to create scatterplots, and how to use scatterplots to answer questions about data. In line with our difficulty factors design approach, our first goal was to develop a first-draft understanding of what skills students needed to learn to create and use scatterplots, and what factors presented difficulty to students learning these skills.

A survey of the relevant research literature determined that although there had been considerable research on students' difficulties in learning to interpret scatterplots, there had been considerably less research on what factors make it difficult for students to learn to create scatterplots (cf. Shah and Hoeffner, 2002). Furthermore, many of the graph creation studies that existed involved graphing abstract linear functions (e.g. Chiu, Kessel, Moschkovich, & Muñoz-Nuñez, 2001), which we thought likely to present different difficulty factors than scatterplots of data. Finally, many of the studies which did focus on scatterplots concerned groups of students creating scatterplots together (cf. Lehrer & Schauble, 2001; Cobb, 2002), rather than on the individual cognition involved in learning to create scatterplots.

In this section, we will present two studies we conducted in order to determine what some of the main difficulty factors are for students learning to create scatterplots. Both of these studies were DFAs.

Difficulty Factors Assessment One (DFA1)

Our first study of the difficulty factors in scatterplot generation (called study DFA1) was part of a Sub-Domain Support DFA given to 52 students in order to investigate their ability to create, interpret, and select between several representations of data, prior to our research group developing a set of cognitive tutor lessons on data representation. In brief, this study determined that graph interpretation problems were considerably easier than graph creation problems, within the representations studied, and that both scatterplot creation and histogram creation were difficult for students.

At this point, we were able to analyze the scatterplot creation and histogram creation portions of DFA1, to determine which errors students made. In this section, we discuss our error analysis of the scatterplot creation portion of study DFA1 – full detail on other aspects of the much broader original study is given in (Baker, Corbett, & Koedinger, 2001).

Participants and Study Design

12 students created scatterplots, in the scatterplot-creation portion of study DFA1. The students were randomly selected from 8th and 9th grade pre-algebra classes in the Pittsburgh suburbs. Half of the students were male and the other half were female. Study DFA1 was conducted prior to the year's data analysis unit; students had some exposure to scatterplots in the prior two years' curricula, although the previous curricula had emphasized scatterplot interpretation rather than creation.

Each student in study DFA1 was given (on paper) a set of data, in the form of a table, shown in Figure 2. This table contained two quantitative variables which could be used to draw the requested representation, plus one nominal variable as a distractor. The student was then asked to create a scatterplot. In order to assess students' conceptual knowledge of what properties a scatterplot had, the students' task did not include contextual information such as giving a student a question to answer, or telling students what variables to use.

Brands of Creamy Peanut Butter	Quality Ratings (1-100)	Price per serving ($)
Captain Hook	61	$0.71
Country Fine	92	$1.09
Davis	54	$0.65
Delish	3	$0.21
GnuMade	61	$0.75
GrubClub	58	$1.09
Jip	21	$0.29
Jolly Eagle	29	$0.47
Kalgan	19	$0.37
L. Butler's	85	$1.06
Old Biff	35	$0.66
Old Georgia	51	$0.68
Scrappy	77	$0.48
Skap	75	$0.94
Tucker's Regular	37	$0.55

Musician	Age	Pieces of fan mail (thousands)
CJ	22	11
Nick	20	11
Devin	23	7
Glennis	25	5
Howie	24	4
Britney	19	8
Ryan	23	4

Fig. 2. The data tables used in study DFA1 (left) and study DFA2 (right)

The task was:

```
Please draw a scatterplot, showing all of the data in this
table. Show all work. Feel free to use graph paper, if necessary.
```

Because we were interested in students' understanding of the concept of a scatterplot, rather than their ability to recognize the word "scatterplot", we gave the students a sheet showing them an example of a scatterplot, as well as examples of the other representations involved in the overall Sub-Domain Selection DFA – histograms, box plots, and stem-and-leaf plots.

Results

None of the twelve students was able to create a completely correct scatterplot or even a graph with appropriate variables and visual appearance. Six of the twelve students simply left the graph blank. One student wrote gibberish on the page, and another copied the box plot from the example sheet (with variables "wingspan" and "kilometers flown"). Two students wrote down the names of the correct variables on the graph but did not proceed further; one student wrote down the names of the correct variables on the graph, but did not label the axes with values, and drew apparently random points.

In total, only two students produced a graph which clearly corresponded to the data in the data table – but this graph was still clearly not a correct scatterplot. These students chose axis variables that were not appropriate for a scatterplot, but were appropriate for a different representation which was not mentioned anywhere in the study or materials, a bar graph (an example is shown in Figure 3). This behavior appeared in students' attempts to create histograms as well (cf. Baker, Corbett, & Koedinger,

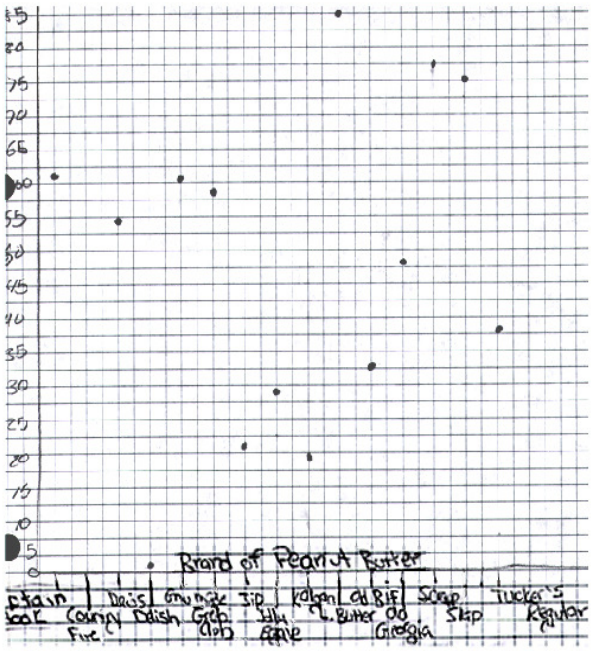


Fig. 3. An example of the variable choice error, in study 1
(image digitally enhanced for readability).

2001). These students chose an x-axis which represented individual values of a nominal variable and a

y-axis which represented values of a quantitative variable, making these graphs the informational equivalent of a bar graph [Footnote 1]. We hypothesized at this point that this behavior could potentially involve over-generalized knowledge rather than complete lack of knowledge, and that this over-generalization could give students difficulty in learning the correct knowledge later (cf. Clement, 1982, 1987). We termed this behavior the *variable choice* error.

While the variable choice error was interesting, the fact that no student created a correct scatterplot was – at the time – considerably more striking. We hypothesized that students might have failed to create correct scatterplots because of the lack of contextual information in the problem situation, and that perhaps if the students had understood *why* they were creating a scatterplot, they would have created correct scatterplots. (Note, however, that any student who fully understood what a scatterplot was would have been able to create a correct scatterplot using the information given; several adults were able to quickly and easily draw appropriate scatterplots given the same materials the students received).

In study DFA2, we attempted to probe both whether contextualizing the problem situation would improve performance, and what factors would affect the variable choice error.

Difficulty Factors Assessment Two (DFA2)

In study DFA2, we investigated in greater detail why students had so much difficulty creating scatterplots. In this study, we gave students a problem similar to that used in DFA1, but varied the scaffolding given on the problem, in order to see which scaffolds affected students' ability to perform the task – and therefore which skills were most difficult for students. In particular, we investigated the possibility that students' low success at creating scatterplots was due to the lack of context in the problem statement. This study also investigated whether specific scaffolds would prevent students from making the variable choice error.

Study DFA2 is an example of a Step Support DFA study; in many cases, it might be reasonable for a Strategy Selection DFA to follow a Sub-Domain Selection DFA. The process of creating a scatterplot, however, has a constrained order (for instance, it is impossible to plot points without first having labeled the axes with values, and in many cases it is impossible to label the axes with appropriate values without having first chosen what variables to place on each axis). Hence, we assumed that correct approaches to creating a scatterplot will follow similar overall procedures, and moved directly to investigating which steps of the scatterplot creation process are particularly difficult for students, by investigating whether offering support on individual steps made the overall process of creating a scatterplot easier.

At the same time as this study, we followed up another sub-domain found to be difficult for students in study DFA1, histogram creation. The Step Selection DFA used to investigate student difficulties in histogram creation is discussed in more extensive detail in (Baker, Corbett, & Koedinger, 2002).

Participants and Study Design

119 students participated in study DFA2. All students were in 8th and 9th grade pre-algebra classes, in inner-city and suburban Pittsburgh. As with study DFA1, study DFA2 was conducted prior to the year's data analysis unit; students had some exposure to scatterplots in the prior two years' curricula, although the previous curricula had not emphasized scatterplot creation. The prior years' curricula did emphasize the creation of bar graphs.

We investigated the effects of two common types of educational scaffolding in this study. First, explicitly telling the students the question they should be able to answer with the representation they

created (which included the names of the variables); second, explicitly writing the name of which variable the student should use alongside the X and/or Y axes. The first type of scaffold is used in most existing middle school curricula that involve scatterplot creation; the second scaffold is used in some reform curricula (e.g. Teaching Integrated Math and Science [TIMS] Project, 1999).

The first scaffold – explicitly telling students what question they should be able to answer – was given to all students. The prompt was as follows:

```
This data shows the ages of several musicians and the number of
pieces of fan mail they receive each day, in thousands.
Please draw a scatterplot, to show if the amount of fan mail a
musician gets is related to their age.
Show all work.
Hint: Scatterplots are made up of dots.
```

We used performance in study DFA1 as an informal comparison condition for this scaffold. Since the problems varied in a few other ways, this comparison is only informal.

The second scaffold – directly labeling one or more axes with the name of the variable to use – was divided into four between-subjects conditions: no-axes-labeled, X-axis-labeled, Y-axis-labeled, and both-axes-labeled. In all four conditions, each student completed an exercise where they were asked to generate a scatterplot from a table with two quantitative variables and one distractor nominal variable.

The data table used in study DFA2 is shown in Figure 2. Beyond the inclusion of the scaffold, fewer data points were included in study DFA2 than DFA1, in order to make it easier to plot all of the points in a short period of time (note, however, that having too many points does not explain the low performance in study DFA1 – no student in study DFA1 plotted any points correctly). The data scales were also made lower (necessitating a change in cover story), in case the relatively large numbers had caused students to give up instead of trying to create the graph (note, however, there was no evidence of students attempting to label the axes with scales but then giving up, in the data from study DFA1).

Results

Giving the students a contextualized question to answer appeared to have a powerful effect on their ability to create scatterplots. As shown in Table 1, 38%-53% of the students in the four conditions of study DFA2 created a completely correct scatterplot. There was not a significant difference in the frequency of completely correct scatterplots between any of the conditions in study DFA2 -- the largest difference between any two conditions was Z=1.19, p=0.24, using the test of the significance of the difference between two independent proportions (Ferguson, 1971, p.160-162). The 38-53% frequency of completely correct scatterplots in study DFA2 was much higher than the 0% frequency in study DFA1 – the difference between study DFA1 and any of the four conditions of study DFA2 was statistically significant – the smallest difference for any condition was Z=2.49, p=0.01, using the test of the significance of the difference between two independent proportions.

The frequency of the variable choice error did not appear to be reduced through just adding a question, but significantly fewer students made the variable choice error if they received a graph with the X axis already labeled. Students given a graph with the X axis labeled (the X-axis-labeled and Both-axes-labeled conditions) made the variable choice error 7% of the time, whereas the remaining students (Y-axis-labeled and No-axes-labeled) made the choice error 24% of the time, $\chi^2(1, N=119)=6.75$, p=0.01. Nonetheless, the continued prevalence of the variable choice error even when the graph was explicitly labeled with a different variable suggested that this error could be a resilient

misconception (cf. Clement, 1982, 1987) that would need to be explicitly addressed within a tutor lesson on this subject.

Table 1
Scatterplot generation in studies DFA1 and DFA2, showing the frequency of each of the common results for each of the different prompts given.

	Variables not given (Study DFA1)	No labels	X axis labeled	Y axis labeled	Both axes labeled
Blank or Uninterpretable Graph	67%	3%	21%	24%	13%
Variable Choice Error	17%	23%	7%	24%	6%
Nominalization Error	n/a	13%	21%	10%	16%
Completely Correct Scatterplot	0%	53%	38%	38%	45%

Interestingly, though many more students chose the correct variables, many went on to make a second error which appeared to be linked to knowledge about bar graphs. Across conditions, 10%-21% of students made what we termed the *nominalization error*, transforming a quantitative variable into a nominal variable, as shown in Figure 4. Instead of plotting the values of a variable in numerical order, with even intervals and with the same value plotted in the same location on the graph, even if it occurred twice (for instance, 19 20 21 22 23 24 25), these students plotted the individual values of the variable, with the same value (23) repeated twice, and in many cases in the same order as in the table: 22,20,**23**,25,24,19,**23**. [Footnote 2] Despite the fact that these students chose the correct variables, the representations these students drew were informationally equivalent to a bar graph, with one nominal variable and one quantitative variable rather than two quantitative variables.

Hence, students learning to create scatterplots demonstrated two distinct errors which nonetheless appear to have a common source: confusing the informational content of scatterplots with the informational content of bar graphs, a different data representation. This can also be viewed as the overgeneralization of knowledge which is appropriate and correct in one context (bar graphs) to a

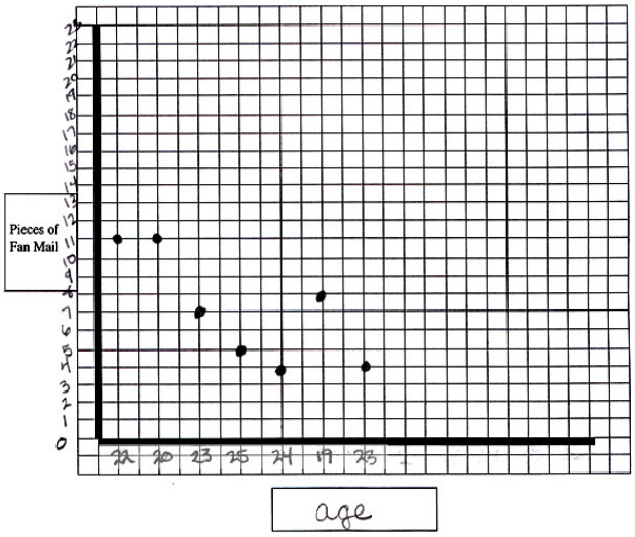


Fig. 4. An example of the nominalization error, in study 2 (image digitally enhanced for readability).

context where it is inappropriate (scatterplots) (cf. Clement, 1982, 1987; Holyoak, 1985). The pattern of behavior demonstrated across these two studies suggests that many of these students had very shallow knowledge about what a scatterplot is, despite having previously studied scatterplots in class. In specific, it appears that these students believe that all (or many) types of graphs contain the same types of information – one quantitative variable and one nominal variable.

PHASE TWO: A FIRST-DRAFT COGNITIVE TUTOR LESSON

After analyzing the data from the two DFA studies, we felt we had reasonable awareness of what some of the main difficulty factors were in this domain, and began developing a first-draft cognitive tutor lesson to help students learn to create scatterplots. Our first step was to see if allowing students to make the variable choice and nominalization errors, in combination with the types of learning support traditionally used in cognitive tutors (such as instant feedback including "bug messages", and context-sensitive help) would be sufficient to help students learn to avoid these errors. In this section, we discuss the design of this first-draft cognitive tutor lesson, and a study (termed study T1) to investigate its effects on student learning.

Prior Systems

An important step towards developing a first-draft tutor lesson was learning how previous educational interfaces (including paper-based curricula) had addressed the creation of scatterplots or similar graphical representations.

One striking commonality to the interfaces we found was that many of them appeared to have been designed (whether intentionally or not) to prevent the student from making the variable choice and nominalization errors.

For instance, Tabletop, a widely used system for teaching students to analyze data with graphs (cf. Hancock, Kaput, & Goldsmith, 1992) prevented students from making the nominalization error. In Tabletop, when the student chooses an axis variable, appropriate upper and lower bounds and scale are automatically chosen for the axis, and values are automatically written along the axis. The variable choice error was also not directly dealt with in Tabletop – whatever variables a student chooses, Tabletop will create a representation to display them, and it is left to the teacher to intervene if the student chooses variables that cannot be used to answer the question of interest.

Cognitive Tutor Algebra (Koedinger, Anderson, Hadley, & Mark, 1997), another widely-used system, tutors the creation of function graphs, a representation similar to scatterplots. Like Tabletop, Cognitive Tutor Algebra does not provide students the opportunity to make the variable choice error, in this case by never having nominal variables available for the student to choose. Similarly, students are not given the chance to express the nominalization error: The student chooses each axis's lower bound, upper bound, and interval between labels – as in the graphing calculators many classrooms use – and the system then labels numbers along the axis for the student.

MathTrailblazers, a paper curriculum for mathematics and science which focuses on data representation (TIMS, 1999), also used this sort of scaffolding. Problems in MathTrailblazers frequently labeled both for the variables to use and the values along the axes – preventing students from making the variable choice and nominalization errors. However, MathTrailblazers also included exercises which allowed students to make these errors, giving the teacher an opportunity to observe and correct these errors.

While two of these interfaces prevent students from making errors in choosing and labeling axes, all three interfaces emphasized a different skill – plotting points. In Tabletop and MathTrailblazers,

students plot every data point, on every problem; in Cognitive Tutor Algebra, students plot every data point on early problems. Hence, all three prior interfaces emphasized a skill which our difficulty factors research suggested is not particularly difficult for students.

Tutor Design

In the light of the evidence from our two DFA studies, we believed that it was important to design a cognitive tutor lesson which gave students the opportunity to make the variable choice and nominalization errors, so that the tutoring software could recognize and respond to these errors. To test this prediction, we built a comparative study into the deployment of our first-draft tutor lesson on creating scatterplots. We created two variations of our tutor lesson, differing in whether they allowed the nominalization error. In both variants, the variable choice error was allowed.

In the tutor lesson, students were given a set of problems. Each problem had a set of variables (including the variables to use, and both quantitative and nominal distractor variables), and a question to answer. The student's task was to create a scatterplot that could be used to answer the given question. In accordance with Anderson et al's (1995) design principles for cognitive tutors, student learning was supported by immediate feedback – if a student made an error, their input turned red, and if the student's action indicated a known bug (such as the variable choice error, and the nominalization error in the tutor condition where it was possible), then the tutor popped up a remedial explanation of why the behavior was incorrect (without giving away the correct answer).

Students created scatterplots within the tutor as follows: first, the student chose and labeled the variables on the X and Y axis, by typing a variable name into a blank along each axis (labels 1 and 2 in Figure 5). A student who chose a nominal variable was given immediate feedback.

Next, the student chose each axis's bounds and scale. In the tutor condition where students could make the nominalization error, the student determined the bounds and scale implicitly, by typing values into slots below the axis (label 3b in Figure 5). In the other condition, the nominalization error was prevented with a scaffold which had the student explicitly choose values for the min, max, and

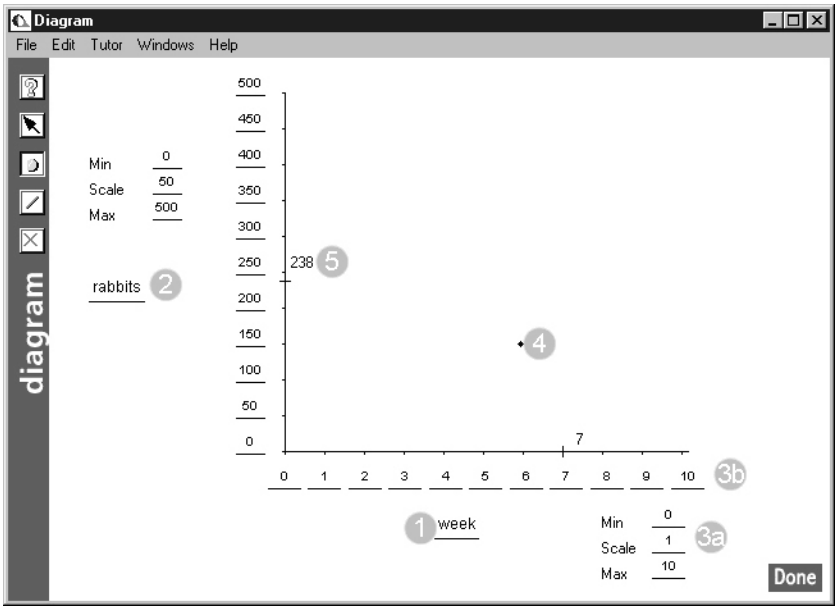


Fig. 5. The tutor user interface in study T1

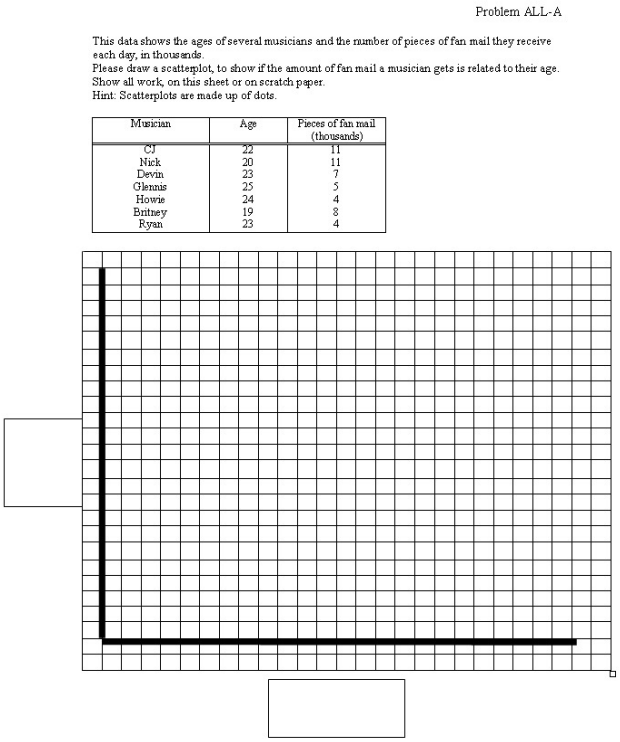
scale (label 3a in Figure 5) – afterwards, the system labeled the values along the axis.

Finally, the student plotted points on the graph by clicking on the point tool and then clicking on the graph where they wished to place the point. (for example at label 4 in Figure 5) A hashmark on each axis (label 5 in Figure 5) indicated the mouse's current location along the axis, to prevent the student from making errors due to not being able to visually translate the cursor's location across the screen.

Participants and Study Design (Study T1)

46 students participated in study T1. All students were in 8th and 9th grade mainstream pre-algebra classes, in inner-city and suburban Pittsburgh. All of the students had just completed a unit of conceptual instruction on data representation in class. Each of the 46 students completed a pre-test, a post-test, and at least four problems in the tutor. Every student used the tutor for the same amount of time, so some students were able to complete more problems than others – the problems were given in the same order for each student. Before students started the tutor, they were shown a practice problem, during which the user interface was explained to them. After this, they worked through the tutor on their own.

Both the pre-test and the post-test consisted of one of two nearly isomorphic problems, counterbalanced between the pre-test and post-test. In each problem, students were given a data table



Problem ALL-A

This data shows the ages of several musicians and the number of pieces of fan mail they receive each day, in thousands.
Please draw a scatterplot, to show if the amount of fan mail a musician gets is related to their age.
Show all work, on this sheet or on scratch paper.
Hint: Scatterplots are made up of dots.

Musician	Age	Pieces of fan mail (thousands)
CJ	22	11
Nick	20	11
Devin	23	7
Glennis	25	5
Howie	24	4
Britney	19	8
Ryan	23	4

Fig. 6. One of the tests used in studies T1 (also the No-axes-labeled condition of study DFA2). This test was pre-test for half of the students, and post-test for the other half. Items in study T2H and T2C were identical, except for the incorporation of an additional quantitative distractor variable in those studies.

with two quantitative variables and one distractor nominal variable. The same problems were given in study T1 as in study DFA2. A question was included, as in study DFA2, but neither of the axes were labeled. Hence, the pretest and posttest were equivalent to the No-axes-labeled condition of study DFA2. A picture of one version of the tests used is shown in Figure 6.

Results (Study T1)

The tutor lesson was effective, across conditions, at improving students' ability to generate scatterplots. Table 2 shows the overall pattern of errors on the pre-test and post-test. The percentage of students who made any type of error decreased significantly from 50% at pre-test to 28% at post-test, Z=2.5, p=0.01, using the test of the significance of the difference between two correlated proportions (Ferguson, 1971, p.162-164). However, the variable choice error's incidence did not reduce significantly, going from 17% on the pre-test to 13% on the post-test, Z=0.82, p=0.41, using the test of the significance of the difference between two correlated proportions. Of the students who committed the variable choice error on the post-test, half had committed it on the pre-test – the other half had not committed it on the pre-test.

Unexpectedly, it was not possible to compare the frequency of the nominalization error between conditions. A small number of students made the nominalization error on the pre-test, but once we excluded students who were not present for the entire study, the study contained no students who had committed the nominalization error. Thus, we were unable to investigate the difference between the two different tutor conditions' effects on the nominalization error.

Another type of student difficulty, in choosing the appropriate scale for a given axis, appeared on both the pre-test (17%) and post-test (15%). As found in (Kerslake, 1981), errors in choosing scales were not consistent across students. Some students chose scales which were too small, making it impossible to put all of the points in the graph. Other students chose scales which were too large, compressing all of the points into a very small area of the graph. It is curious that scaling errors did not appear in any significant quantity until this point (across conditions, 4% of students appeared to make scaling errors in study DFA2). It is possible that in studies DFA1 and DFA2, scaling errors were much rarer because students who would have made scaling errors made variable choice or nominalization errors first, preventing scaling errors.

Hence, the tutor's overall effects on student performance were positive, but it was also clear from this study that the tutor still required considerable improvement. In particular, just allowing students to make the variable choice error and then addressing it with the instant feedback, bug messages, and context-sensitive help characteristic to cognitive tutors, was not sufficient to assist those students in learning to choose appropriate variables, a key part of learning to create scatterplots.

Table 2
Percentage of students making each error in study T1, on pre-test and post-test

Error	Pre-Test	Post-Test
Variable Choice Error	17%	13%
Nominalization Error	0%	0%
Point Plotting Error (>1 point incorrect)	9%	0%
Scaling Errors	17%	15%
General Miscomprehension Errors	7%	0%

PHASE 3: TUTOR REDESIGN

Our two paper studies and first tutor study showed that the variable choice error is resilient to simple forms of scaffolding, such as telling the student what variables to use, or giving feedback as to why the student should not place a nominal variable on the X axis. We decided to emphasize what data goes into different representations within the next version of the tutor, creating a scaffold to help students learn the conceptual differences between scatterplots and bar graphs. We also developed a scaffold to help students learn to choose an appropriate scale and bounds for each axis.

In this section, we discuss the design of these scaffolds, and a pair of studies we conducted to test their effectiveness, in two very different settings: with homeschool students via distance-learning, and in suburban schools.

Scaffold Design

A Contrasting Cases Scaffold

The largest challenge in redesigning our tutor lesson was to determine how to help students learn to avoid the variable choice error. We had already determined that this error was resilient to the sorts of scaffolding used in some prior curricula (in study DFA2), and to the sort of scaffolding typically used in cognitive tutors (in study T1). The nature of the variable choice error – and the sporadic presence of the nominalization error – suggested that the variable choice error might arise from a failure to distinguish between the information contained in scatterplots and bar graphs, an overgeneralization (cf. Clement, 1982, 1987; Holyoak, 1985) of knowledge from bar graphs to scatterplots. Therefore, our primary goal in developing a scaffold was to help students learn this distinction – helping the student make sense of when and why he or she should create a graph with two quantitative variables.

To this end, we settled on an approach that built upon on prior examples of scaffold design in Schwartz and Bransford (1998), and Tennyson and Cocchiarella (1986). In both of these prior designs, students learned by comparing between cases which differed in specific important features. Contrasting cases, both sets of authors proposed, assist students in learning which characteristics are important to attend to in distinguishing between categories. Both sets of authors found that contrasting cases are most effective when combined with direct instruction on why the features that differentiate the cases are relevant. In Schwartz and Bransford (1998), contrasting cases were given first, with conceptual instruction drawing on the distinctions students learned during the contrasting cases. Tennyson and Cocchiarella utilized the reverse order, and gave conceptual instruction first.

In our design, we adopted Tennyson and Cocchiarella's ordering, which offered several advantages for our instructional context: First, this ordering enabled us to combine conceptual instruction on the features differentiating scatterplots from bar graphs (i.e. the conceptual instruction for the contrasting cases) with conceptual instruction on how to generate a scatterplot – so that students learn what a scatterplot is at the same time they learn what distinguishes it from other representations. Secondly, this ordering enabled us to integrate the contrasting cases into the process of creating a scatterplot used to answer specific questions, with every problem demonstrating the practical utility of the distinction between the informational content of bar graphs and scatterplots.

The contrasting cases were integrated into a scaffold inspired by the data format scaffold in Lovett's Statistics Tutor (2001). As part of helping students learn to choose what kind of representation to use to answer a question, students using the Statistics Tutor are guided to identify each variable's type (within the data table) and to select a representation appropriate to the variables in the question. In order to remediate the variable choice error, we added an explicit contrast of the suitability of each variable for each representation, based on the type of information contained in that variable and the type of information used in the representation.

Our contrasting cases scaffold is shown in Figure 7. In this scaffold, each variable in the data set is listed, and for each variable the student must first identify whether it is a quantitative ("numerical") variable or a categorical variable. After doing so, the student must identify whether that variable is

appropriate or inappropriate for a scatterplot (quantitative variables are appropriate, categorical variables are not), and whether that variable is appropriate or inappropriate for a bar graph (a bar graph

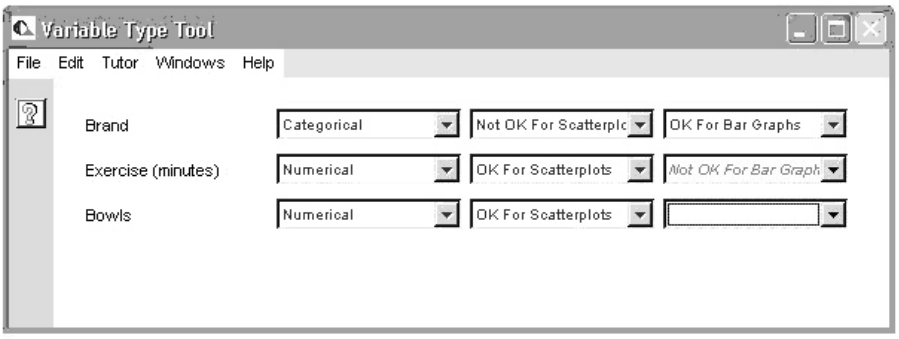


Fig. 7. The contrasting cases scaffold used in studies T1 and T2

uses one variable of each type, so taken individually, a variable of either type is appropriate for use in a bar graph).

By having the student decide whether each variable would be appropriate for each representation, the scaffold assists the student in understanding the distinctions between the types of information used in these two representations of data. Moreover, the student makes this distinction immediately after considering the feature (variable type) that distinguishes the cases, reinforcing the connection between the contrasting cases and the feature that contrasts them.

An Expert Scale-and-Bounds Scaffold

The evidence from study T1 suggested that another skill that was difficult for students to acquire was choosing appropriate bounds and scale. In the tutor in study T1, students learned to choose bounds and scale by selecting these values, with instant feedback and context-sensitive help. However, that support was not sufficient to assist students in learning these skills. In study T1, students made a variety of errors when choosing bounds and scale, both when using the tutor and on the post-test. The variety of errors suggested that although these skills were important difficulty factors, the difficulties for students did not stem from a single overriding misconception.

Our first step towards addressing this problem was to analyze this aspect of our first-draft tutor interface in terms of the existing design principles for intelligent tutors (Anderson et al, 1995), using the method of Heuristic Evaluation (Nielsen, 1994). In doing this, it was rapidly apparent that our design did not succeed in following Anderson et al's principle: "Communicate the goal structure underlying the problem-solving". By simply asking the student to enter bounds and a scale, our system adequately represented the overall process of creating a graph but did not communicate the goal structure of the sub-process of selecting bounds and a scale. Of course, before study T1, we did not realize we *needed* to communicate the goal structure of this sub-process – which is why the analysis of difficulty factors needs to continue throughout the entire process of design.

In order to figure out the best way to represent this sub-process, we investigated how experts (at creating graphs) represent the sub-process to themselves (cf. Lovett, 1998). We asked a number of graduate students in our department, each of whom was able to quickly draw an appropriate scatterplot, to assist us in this, by thinking aloud while drawing scatterplots (so that we could study their processes), and engaging in participatory design (cf. Beyer and Holtzblatt, 1998) with the first author. It is worth noting that this process of studying expert strategies can be seen in some ways as a substitute for the second type of DFA study discussed earlier in this paper, Strategy Selection DFAs.

Though multiple strategies for selecting scale and bounds are possible, in this case, we decided to study experts instead of conducting a DFA, because we did not yet have a satisfactory model of how experts conduct this fairly complex process. In cases where expert performance is better understood, it may be appropriate to conduct a Strategy Selection DFA instead, in order to determine whether some student strategies may be more effective than expert strategies or may serve as a bridge to expert strategies. The participatory design resulted in the development of a scaffold for selecting scale and bounds, shown in Figure 8.

A student using this scaffold determines, for each variable, the variable's range. The student then goes through an iterative process of choosing a scale and evaluating whether that scale is appropriate (neither squeezing all data into a small space nor placing some data outside of the graph). The student finally chooses the first label for the axis – a "round" number just below the minimum value of the data set.

Interestingly, this process does not involve explicitly choosing a maximum value for the axis. Instead, the axis is labeled using the chosen first label and scale, and the student stops once the maximum value of the data set has been surpassed. This contrasts to the interface used in study T1, the interface used in Cognitive Tutor Algebra (Koedinger et al, 1997), and the interface used in most graphing calculators. While selecting a maximum value is a perfectly reasonable part of an interface for creating graphs in an automated fashion, it does not appear to be part of the process experts use in the absence of a graphing calculator interface.

Homeschool Study (Study T2H)

We conducted two studies to investigate the educational effectiveness of our scaffolds and re-designed tutor lesson. The first study (Study T2H) was in a homeschool setting; the second study was in a classroom setting (Study T2C). The two studies occurred concurrently.

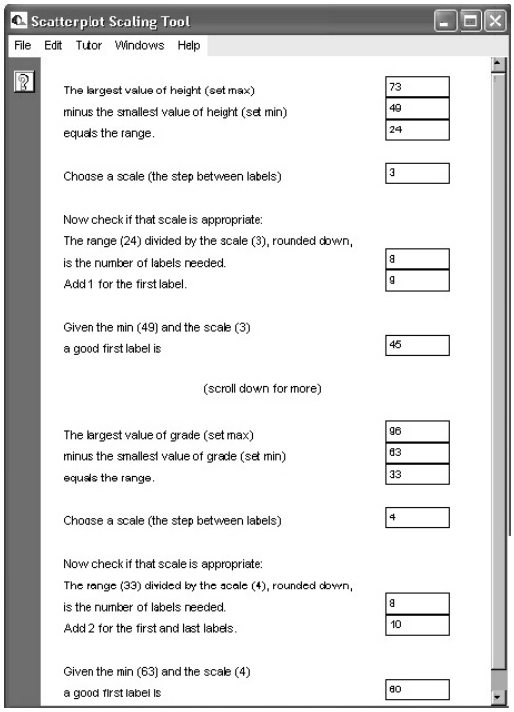


Fig. 8. The expert scale-and-bounds scaffold

Conducting a study with homeschool students allowed us to assign students randomly to different conditions without the risk of students discussing the differences in their tutoring software (and thus contaminating a comparison between conditions). Hence, in study T2H, we focused on comparing the effects of having or not having each of the scaffolds.

Students in Study T2H used tutors which varied along two dimensions. Half of the students received a cognitive tutor which included the contrasting cases scaffold and conceptual instruction designed to be used in combination with this scaffold (condition CONTRASTING-CASES). The other half of students received a cognitive tutor and conceptual instruction which taught about scatterplots, but which did not explicitly focus on the differences between scatterplots and bar graphs (condition SCATTERPLOT-ONLY). An orthogonal comparison was also made, where half of students received a cognitive tutor which included the expert scale-and-bounds scaffold (condition SCALING-SCAFFOLD) and the other half received a cognitive tutor where scale and bounds were chosen by simply indicating scale, min, and max, as in our previous tutor and graphing calculators (condition SCALING-TRADITIONAL). Conceptual instruction in each condition corresponded to the method used to select scale and bounds in the tutor.

Conceptual instruction was given in all conditions via a PowerPoint presentation with voiceover and simple animations. Students went through the PowerPoint presentation at their own pace, although the presence of voiceover tended to keep the students to approximately the same total time. The instruction was designed to provide students with basic conceptual knowledge on how to create and interpret scatterplots, and on what scatterplots could be used for. The majority of the instruction was identical across conditions, varying only when directly appropriate to the difference between conditions. Across conditions, the instruction was of comparable length.

Participants and Study Design (Study T2H)

39 homeschool students in the United States between 11-13 years old participated in Study T2H. Homeschool students were recruited by posting ads on homeschooling newsgroups and internet mailing lists, and were sent all materials via mail. Participating students took the pre-test (administered by their parents), viewed a PowerPoint presentation giving conceptual instruction, used the cognitive tutor, took the post-test, and finally returned the tests and tutor log files to us by self-addressed stamped envelope.

The students used the software on their own computers, at home; we requested that the students' parents and siblings not interact with them as they used the software. If two children from a single family used the software, they were placed in the same condition; families with multiple children were distributed randomly between conditions.

We controlled the amount of time each student spent on the tutor. The tutor allowed students to work for 75 minutes; after that time, the student was allowed to complete the current problem, and then the system quit, telling the student they had completed their work with the tutor. A sufficient number of problems were included in the tutor that no student ran out of problems to work on.

Both the pre-test and the post-test consisted of one of two nearly isomorphic problems, counterbalanced between the pre-test and post-test. In each problem, students were given a data table with two quantitative variables one distractor nominal variable, and one distractor quantitative variable. The problems were highly similar to the items in study T1; the only difference was that one more distractor variable (quantitative) was present in this study's assessments than were present in study T1.

Results (Study T2H)

Average performance improved substantially from pre-test to post-test, across conditions, $F(1,38)=40.65$, $p<0.001$. In the SCATTERPLOT-ONLY condition, average performance improved considerably, from 44% to 88%, but performance improved significantly more in the CONTRASTING-CASES condition, from 43% to 99%, $F(1,36)=6.41, p=0.02$. Hence, the focus on the differences between bar graphs and scatterplots seems to have been educationally beneficial.

The CONTRASTING-CASES condition, as well as producing higher overall gains, was effective at remediating the variable choice and nominalization errors. The variable choice error occurred 22% of the time at pre-test in the CONTRASTING-CASES condition, and 0% of the time at post-test, $Z=2.24$, $p=0.03$, using the test of the significance of the difference between two correlated proportions. Similarly, the nominalization error occurred 13% of the time at pre-test in the CONTRASTING-CASES condition, and 0% of the time at post-test, $Z=1.73$, $p=0.08$, using the test of the significance of the difference between two correlated proportions. Direct comparison to the SCATTERPLOT-ONLY condition was unfortunately not possible, however, since students randomly assigned to that condition unexpectedly did not make either error on the pre-test in that condition. Data from the comparison of the CONTRASTING-CASES and SCATTERPLOT-ONLY conditions is shown in Table 3.

It was not possible to compare between the SCALING-SCAFFOLD and SCALING-TRADITIONAL conditions, due to bad luck in the random assignment of students into conditions. Students in the SCALING-SCAFFOLD condition had much greater success at choosing appropriate bounds and scale at pre-test (the average student selected correct bounds and scale on 1.12 axes) than students in the SCALING-TRADITIONAL condition (where students successfully selected correct bounds for 0.62 axes). Average pre-test scores also differed substantially (43% in the SCALING-SCAFFOLD condition and 29% in the SCALING- TRADITIONAL condition).

Classroom Study (Study T2C)

A second study (T2C), conducted during the same period of time as T2H, one year after study T1, investigated the effectiveness of the two scaffolds in a classroom setting. Within study T2C, we tested a system for focusing each student's time on the parts of the problem which that individual student found most difficult – that system, discussed in more detail in (Baker, Corbett, & Koedinger, 2004), did not have a significant effect on learning. Additionally, study T2C was overlaid with a non-invasive observational study of student off-task behavior, discussed in detail in (Baker, Corbett, Koedinger, & Wagner, 2004). The observations did not affect the content or presentation of the tutoring software, and the observers did not interact with the students.

Beyond these side studies, Study T2C was intended to serve as a test of whether the contrasting cases and expert scale-and-bounds scaffolds were effective in a classroom setting. All students in this study used those scaffolds.

Table 3
Percentage of students making the variable choice and nominalization errors in study T2H, on pre-test and post-test

	SCATTERPLOT-ONLY		CONTRASTING-CASES	
	Pre-Test	Post-Test	Pre-Test	Post-Test
Average Overall Correctness	44%	88%	43%	99%

Variable Choice Error	0%	0%	22%	0%
Nominalization Error	0%	0%	13%	0%

Participants and Study Design (Study T2C)

77 students participated in Study T2C. All students were in 8th and 9th grade mainstream pre-algebra classes, in the Pittsburgh suburbs. All of the students had just completed a unit of conceptual instruction on data representation, using the same PowerPoint conceptual instruction given to the homeschool students. Each of the 77 students completed a pre-test, a post-test, and at least four exercises in the tutor. The pre-test and post-test were identical to the pre-test and post-test in study T2H. Every student used the tutor for the same amount of time (just under two class periods), so some students were able to complete more problems than others – the problems were given in the same order for each student.

Results (Study T2C)

As in study T1 and study T2H, this tutor lesson was effective at improving students' ability to generate scatterplots: average performance rose from pre-test to post-test, 40% to 71%, $F(1,68)=7.59, p<0.01$. The prevalence of the variable choice error decreased from 19% to 6%, which was significant, $Z=2.32$, $p=0.02$, using the test of the significance of the difference between two correlated proportions. By comparison, in study T1, where students did not use the contrasting cases scaffold and corresponding conceptual instruction, the prevalence of this error did not significantly decrease, going from 17% to 13%. The difference between these two studies was not significant, $Z=1.07$, two-tailed $p=0.29$ (procedure used is drawn from Rosenthal and Rosnow, 1991, p.493). Hence, it is not clear that the modified tutor with the contrasting cases scaffold is significantly better than the tutor in T1, but it is clear that this tutor did lead to a significant reduction in the variable choice error. The nominalization error never occurred on the pre-test in study T2C.

On the other hand, errors in bounds and scale did not appear to decrease from pre-test (19%) to post-test (21%). This may have been in part because many students made errors on the pre-test before reaching this part of the problem, making it impossible to demonstrate their lack of knowledge at choosing bounds and scale. Among the students who succeeded on earlier parts of the problem (and thus could have made errors in bounds and scale), 35% made errors in bounds and scale on the pre-test and 24% made errors on the post-test – an apparent drop, but at best a small one (calculating significance is difficult in this case, since neither full independence nor full non-independence can be validly assumed). Hence, the scaling scaffold does not appear to have substantially improved students' learning of these skills. One potential hypothesis for why the scaling scaffold failed to improve learning may be that many students gamed the system while using the scaffold. Gaming the system is behavior aimed at performing well in an educational task by systematically taking advantage of properties and regularities in the system used to complete that task, rather than by thinking about the material. Students who gamed the system (according to the observers' observations – cf. Baker, Corbett, Koedinger, & Wagner, 2004) did approximately equally well on the pre-test at choosing appropriate scale and bounds as students who did not game the system, $Z=0.93$, $p=0.35$, using the test of the significance of the difference between two independent proportions. However, when we look at the same group of students, on the post-test, (not all students reached the point of being able to choose bounds and scale, on either test – it is only possible to plot appropriate bounds and scale if a quantitative variable was chosen), we find that 5 of the 9 gaming students who reached this point in the post-test made errors in choosing bounds and scale on the post-test, whereas only 1 of the 28 non-gaming students made errors in choosing bounds and scale on the post-test, $Z=3.68$, $p<0.001$, using the test of the significance of the difference between two independent proportions.

Hence, the expert scaling-and-bounds scaffold's lack of effectiveness is likely to be related to students' choice to game the system. There is some evidence that students who game dislike the tutoring software (Baker, Roll, Corbett, & Koedinger, 2005) – and many students informally commented that they particularly disliked the guess-and-test component of the scaffold. Alternatively, it may be that the scaffold's complexity and difficulty led some students to game it instead of expending the effort to learn from it.

To summarize, the evidence from this study – and from the homeschool study – suggests that the contrasting cases scaffold (when used in combination with appropriate conceptual instruction) is sufficient to remediate the variable choice error. The expert scaling-and-bounds scaffold, however, still needs improvement, though potentially because of motivational rather than cognitive reasons.

Table 4
Percentage of students making specific errors in study T2C, on pre-test and post-test

Error	Pre-Test	Post-Test
Average Overall Correctness	40%	71%
Variable Choice Error	19%	6%
Nominalization Error	0%	1%
Scaling Errors	19%	21%

At this point, we are at a reasonable place to conclude our investigation of difficulty factors in scatterplot creation. Through our set of studies, we have learned what factors create difficulty for students learning to generate scatterplots, and have used this knowledge to develop a tutor lesson which addresses all of these difficulty factors. We can have reasonable confidence that no more difficulty factors exist in the domain, because the last two studies addressed all existing difficulty factors and found no new difficulty factors, despite being carried out in very different populations and educational settings. Though the tutor lesson developed does not yet lead to perfect learning in all students, the flaws still remaining in the tutor lesson appear to not be related to the students' cognitive difficulty with the material. Instead, these flaws appear to be due to motivational issues with the material's presentation; the resolution of these motivational issues is outside the scope of this paper, and is discussed in (Baker et al, 2006).

DISCUSSION AND CONCLUSIONS

In this paper, we present a process which enabled us to develop a cognitive tutor lesson on scatterplot creation which effectively addresses the principal difficulty factors in that domain. A significant part of the process of developing that tutor lesson was determining what the principal difficulty factors were. It was only possible to discover what the principal student difficulty factors were by investigating difficulty factors throughout the entire process of design: the variable choice error initially obscured the nominalization error, and both of those errors initially obscured student difficulties in selecting scales and bounds.

Our approach focuses on developing effective learning environments by determining which skills and concepts are most difficult for students to learn, and equally importantly, why these skills and concepts are difficult. The task of determining difficulty factors does not occur only once, at the beginning of design, but continues throughout the process of developing and refining the tutor lesson.

Designing lessons within an intelligent tutor or other type of learning environments by researching difficulty factors has two major benefits. First, it enables us to design for the correct student need.

Prior educational systems and paper curricula in this domain focused considerable student time on plotting points, a skill that did not seem to be difficult for students in any of the populations we investigated. At the same time, these prior systems did not directly address the skills of choosing appropriate variables or (in Tabletop's case) choosing a scale and bounds. The research presented here suggests that these are the skills students most need support for. By explicitly researching difficulty factors, we were able to design a cognitive tutor lesson that provided educational support on the parts of the problem where it was most needed.

Equally importantly, using the difficulty factors approach meant that the researchers' time and effort were not spent needlessly on developing, refining, and evaluating scaffolds for skills that students do not find difficult, such as point plotting. It is relatively easy and quick to develop, administer, and analyze multiple difficulty factors assessments early in the design process. By contrast, designing, developing, and assessing software scaffolds is considerably more time consuming. In particular, designing a system and planning its evaluation without first investigating what factors students find difficult virtually guarantees that either the wrong scaffolds and parts of the problem solving process will be evaluated, or that considerable time will be wasted through evaluating all of the scaffolds in a system, regardless of whether evaluation is needed. Hence, a small amount of additional time spent early in the design process saves substantial time later in the design process.

In addition, the set of skills required to develop and analyze DFAs is generally the same set of skills required to design learning material and evaluate a learning environment: knowledge of the domain, and basic data analysis skill. Hence, DFAs can be used to improve design without needing to add new specialized skills to the design team.

Second, the difficulty factors approach allows us to not just develop more effective tutor lessons, but also to create widely-usable knowledge about the domain we are investigating. The program of research described here both resulted in a more effective tutor lesson and contributed to wider knowledge about student learning of graphs. In general, varying our student populations across the design cycle of this cognitive tutor lesson was an important part of learning how general the difficulty factors we discovered are, and making the knowledge gained through this design process make a broader scientific contribution. One discovery was that the variable choice error appears to be common across populations, and is therefore likely to be inherent to the domain, whereas the nominalization error is considerably more volatile, varying considerably in frequency between populations. This finding suggests that the nominalization error may be an artifact of specific instruction given to students in the past.

Hence, the findings obtained by using the difficulty factors approach can be interesting and applicable beyond just the design of a single system. The reverse is also true. The difficulty factors approach to design is best carried out with continual consideration of how the difficulty factors discovered relate to broader learning theory and the existing scientific literature in the domain. When a difficulty factor is discovered, learning theory can provide a way to understand why that difficulty factor occurs, and can point the way to resolving it. To give an example from the program of research discussed in this paper, we were better able to understand the variable choice error by considering the error's relationship to the concept of overgeneralization in the research literature, and we based our design that addressed this difficulty factor on prior research on how to help students make correct conceptual distinctions (part of which is avoiding overgeneralization). In general, a program of design based upon the difficulty factors approach should not stand outside the main flow of scientific discovery in the learning sciences, but should draw from that literature, and attempt to contribute to it as well.

Another issue worth discussing is the relationship between the difficulty factors approach and the student modeling/adaptive systems literature. In this paper, we have focused mostly on discovering what skills and concepts students find most difficult, in general. One of the goals of adaptive systems such as intelligent tutoring systems is to determine which skills each student finds difficult, in order to

offer tailored support to each student. The degree of variability in the pre-test frequency of the errors we studied, across populations, shows that even if a system is designed with scaffolding for the correct difficulty factors, the system should still adapt to each student's particular difficulty factors (and perhaps each population's particular difficulty factors). In general, a system which assesses each student's knowledge can determine which scaffolds to give a student and which scaffolds can "fade away" (cf. Guzdial, 1995; Baker, Corbett, and Koedinger, 2004) after the student has reached mastery. The difficulty factors approach to design is not a substitute for student modeling; instead, we view it as a way to focus the design of systems that use student modeling.

At essence, the difficulty factors approach is a way to focus design. Focusing on difficulty factors enables the efficient creation of learning environments targeted towards helping students learn the skills they need to learn, not the skills that our intuition might tell us they need to learn. Through focusing on difficulty factors throughout the entire process of designing a tutor lesson on scatterplots, we were able to develop both an effective tutor lesson for this domain, and a considerable amount of generalizable knowledge that should prove useful to later research within this domain. By focusing on what the difficult concepts and skills are, and why these difficult concepts and skills *are* difficult, we can create learning environments which support the right skills, and which are fully centered on the needs of the student.

ACKNOWLEDGEMENTS

This research was supported by an NDSEG (National Defense Science and Engineering Graduate) Fellowship, by a research contract from Carnegie Learning Inc: "Cognitive Tutors for Middle School Mathematics", and by IERI grant number REC-043779 to " Learning-Oriented Dialogue in Cognitive Tutors: Towards a Scalable Solution to Performance Orientation ".

We would like to thank Angela Wagner, Jay Raspat, Meghan Naim, Katy Getman, Pat Battaglia, John Argentieri, and Geoff Kaufman for assisting in the administration of the studies discussed here. We would also like to thank Vincent Aleven, Aaron Bauer, Matthew Easterday, Shelley Evenson, Dave Holstius, Brian Junker, Andy Ko, John Kowalski, Marsha Lovett, Santosh Mathan, Tom Murray, Bethany Rittle-Johnson, Rhiannon Weaver, Jack Zaientz, and the anonymous reviewers for helpful discussions and suggestions.

REFERENCES

Aleven, V., & Koedinger, K.R. (2002) An effective meta-cognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science, 26,* 147-179.

Aleven, V., Koedinger, K.R., Sinclair, H.C., & Snyder, J. (1998) Combatting shallow learning in a tutor for geometry problem solving. *Proceedings of the 4th International Conference on Intelligent Tutoring Systems,* 364-373.

Anderson, J.R., Corbett, A.T., Koedinger, K.R., & Pelletier, R. (1995) Cognitive Tutors: Lessons Learned. *The Journal of the Learning Sciences,* 4, 167-207.

Anderson, J.R., & Lebiere, C. (1998) *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Baker R.S., Corbett A.T., & Koedinger K.R. (2001) Toward a Model of Learning Data Representations. *Proceedings of the 23rd Conference of the Cognitive Science Society*, 45-50.

Baker, R.S., Corbett, A.T., & Koedinger, K.R. (2002) The Resilience of Overgeneralization of Knowledge about Data Representations. Presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, USA.

Baker, R.S., Corbett, A.T., & Koedinger, K.R. (2004) Learning to Distinguish Between Representations of Data: A Cognitive Tutor That Uses Contrasting Cases. *Proceedings of the Fifth International Conference of the Learning Sciences,* 58-65.

Baker, R.S., Corbett, A.T., Koedinger, K.R., & Wagner, A.Z. (2004) Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game The System". *Proceedings of ACM CHI 2004: Computer-Human Interaction*, 383-390.

Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., Evenson, E., Roll, I., Wagner, A.Z., Naim, M., Raspat, J., Baker, D.J., & Beck, J. (2006) Adapting to When Students Game an Intelligent Tutoring System. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 392-401.

Baker, R.S., Roll, I., Corbett, A.T., & Koedinger, K.R. (2005) *Do Performance Goals Lead Students to Game the System?* Proceedings of the International Conference on Artificial Intelligence and Education, 57-64.

Baumgartner, E., & Bell, P. (2002) What Will We Do With Design Principles? Design Principles and Principled Design Practice. Presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, USA.

Beyer, H., & Holzblatt, K. (1998) *Contextual Design: Defining Customer-Centered Systems*. San Francisco: Morgan Kaufmann.

Chiu, M.M., Kessel, C., Moschkovich, J., & Muñoz-Nuñez, A. (2001) Learning to Graph Linear Functions: A Case Study of Conceptual Change. *Cognition and Instruction,* 19 (2), 215-252.

Clement, J. (1982) Students' Preconceptions in Introductory Mechanics. *American Journal of Physics,* 50 (1), 67-71.

Clement, J. (1987) The use of analogies and anchoring intuitions to remediate misconceptions in mechanics. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

Cobb, P. (2002) Reasoning With Tools and Inscriptions. *Journal of the Learning Sciences,* 11 (2&3), 187-215.

Cohen, S., Smith, G., Chechile, R.A, Burns, G., & Tsai, F. (1996) Identifying Impediments to Learning Probability and Statistics From an Assessment of Instructional Software. *Journal of Educational and Behavioral Statistics,* 21 (1), 35-54.

Collins, A., Joseph, D. & Bielaczyc, K. (2004) Design research: Theoretical and methodological issues. *Journal of the Learning Sciences, 13* (1), 15–42.

de Jong, T., & van Joolingen, W.R. (1998) Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research,* 68, 179-202.

Ericsson, K.A., & Simon, H.A. (1993) *Protocol Analysis.* Cambridge, MA: MIT Press.

Guzdial, M. (1995) Software-Realized Scaffolding to Facilitate Programming for Science Learning. *Interactive Learning Environments,* 4 (1), 1-44.

Ferguson, G. (1971) *Statistical Analysis in Psychology & Education.* New York: McGraw-Hill.

Hancock, C., Kaput, J.J., & Goldsmith, L.T. (1992) Authentic Inquiry With Data: Critical Barriers to Classroom Implementation. *Educational Psychologist*, 27(3), 337-364.

Heffernan, N., & Koedinger, K. R. (1998). A Developmental Model For Algebra Symbolization: The Results of a Difficulty Factors Assessment. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, 484-489. Mahwah,NJ: Erlbaum.

Heffernan, N.T. *Intelligent Tutoring Systems have Forgotten the Tutor: Adding a Cognitive Model of Human Tutors.* Doctoral dissertation, published as Carnegie Mellon University School of Computer Science Technical Report CMU-CS-01-127.

Holyoak, K. (1985) The Pragmatics of Analogical Transfer. *Psychology of Learning and Motivation*, 19, 59-87.

Kelley, T. (2001) *The Art of Innovation: Lessons in Creativity from IDEO, America's Leading Design Firm.* New York: Currency Books.

Kerslake, D. (1981) Graphs. In K.M. Hart (Ed.), *Children's Understanding of Mathematics: 11-16,* 120-136. London: John Murray.

Koedinger, K. R. (2002). Toward evidence for instructional design principles: Examples from Cognitive Tutor Math 6. Invited paper in *Proceedings of PME-NA XXXIII (the North American Chapter of the International Group for the Psychology of Mathematics Education)*.

Koedinger, K.R., Aleven, A., Heffernan, N., McLaren, B., & Hockenberry, M. (2004) Opening the Door to Non-Programmers: Authoring Intelligent Tutor Behavior by Demonstration. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems,* 162-173.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent Tutoring Goes to School in the Big City. *International Journal of Artificial Intelligence in Education*, 8, 30-43.

Koedinger, K.R., & Corbett, A.T. (2006) Cognitive Tutors: Technology bringing learning science to the classroom. In K.Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences,* 61-78. Cambridge, UK: Cambridge University Press.

Koedinger, K.R., & Nathan, M.J. (2004) The real story behind story problems: Effects of representation on quantitative reasoning. *The Journal of the Learning Sciences, 13 (2),* 129-164.

Lehrer, R., & Schauble, L. (2001). *Investigating real data in the classroom: Expanding children's understanding of math and science*. New York: Teachers College Press.

Linn, M.C. (1995) Designing Computer Learning Environments For Engineering and Computer Science: The Scaffolded Knowledge Integration Framework. *Journal of Science Education and Technology,* 4 (2), 103-126.

Lovett, M. C. (1998). Cognitive task analysis in service of intelligent tutoring systems design: A case study in statistics. In B. P. Goettl, H. M. Halff, C. L. Redfield, & V. J. Shute (Eds.) *Intelligent Tutoring Systems, Lecture Notes in Computer Science,* Volume 1452, 234-243. New York: Springer.

Lovett, M. (2001) A Collaborative Convergence on Studying Reasoning Processes: A Case Study in Statistics. In D. Klahr & S. Carver (Eds.) *Cognition and Instruction: 25 Years of Progress.* Mahwah, NJ: Erlbaum.

McConnell, S. (1996) *Rapid Development, Taming Wild Software Schedules.* Redmond, WA: Microsoft Press.

Merrill, D.C., & Reiser, B.J. (1993) Scaffolding the Acquisition of Complex Skills With Reasoning-Congruent Learning Environments. *Proceedings of the Workshop in Graphical Representations, Reasoning, and Communication from the World Conference on Artificial Intelligence in Education (AI-ED '93,* 9-16.

Murray, T., Woolf, B., & Marshall, D. (2004) Lessons Learned from Authoring for Inquiry Learning: A Tale of Authoring Tool Evolution. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems,* 197-206.

Nathan, M., & Koedinger, K. (2000) An investigation of teachers' beliefs of students' algebra development. *Cognition and Instruction*, 18(2), 209-237.

National Council of Teachers of Mathematics (2000) *Principles and Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.

Nielsen, J. (1994) *Usability Engineering.* San Diego, CA: Academic Press.

Quintana, C., Reiser, B.J., Davis, E.A., Krajcik, J., Fretz, E., Duncan, R.G., Kyza, E., Edelson, D., & Soloway, E. (2004) A Scaffolding Design Framework for Software to Support Science Inquiry. *The Journal of the Learning Sciences,* 13 (3), 337-386.

Rittle-Johnson, B., & Koedinger, K.R. (2001) Using Cognitive Models to Guide Instructional Design: The Case of Fraction Division. *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society,* 857-862.

Rittle-Johnson, B., & Koedinger, K.R. (2005) Designing better learning environments: Knowledge scaffolding supports mathematical problem solving. *Cognition and Instruction, 23 (3),* 313-349.

Rosenthal, R., & Rosnow, R.L. (1991) *Essentials of Behavioral Research: Methods and Data Analysis.* Boston: McGraw-Hill.

Shah, P., & Hoefnner, J. (2002) Review of Graph Comprehension Research:
Implications for Instruction. *Educational Psychology Review*, 14(1), 47-69.

Siegler, R. (1976) Three Aspects of Cognitive Development. *Cognitive Psychology,* 8, 481-520.

Siegler, R.S. (1987) The Perils of Averaging Data Over Strategies: An Example From Children's Addition. *Journal of Experimental Psychology: General,* 116 (3), 250-264.

Simon, H. (1996) *The Sciences of the Artificial.* Cambridge, MA: MIT Press.

Singley, M.K., & Anderson, J.R. (1989) *The Transfer of Cognitive Skill*. Cambridge, MA: Harvard University.

Teaching Integrated Math and Science (TIMS) Project. (1999) *Math Trailblazers*. Dubuque, IA: Kendall/Hunt Publishing Company.

Tennyson, R.D., & Cocchiarella, M.J. (1986) An Empirically Based Instructional Design Theory for Teaching Concepts. *Review of Educational Research,* 56 (1), 40-71.

van Lehn, K. (1990) *Mind Bugs: The Origins of Procedural Misconceptions*. Cambridge, MA: MIT Press.

van Lehn, K., Lynch, C., Schulze, K., Shapiro, J.A., Shelby, R., Treacy, D., Weinstein, A., & Wintersgill, M. (2005) The Andes Physics Tutoring System: Lessons Learned. *International Journal of Artificial Intelligence in Education,* 15, 147-204.

Vygotsky, L.S. (1978) *Mind in Society: The Development of Higher Psychological Processes.* Cambridge, MA: Harvard University Press.

FOOTNOTES

Footnote 1: Although “bar graph” can be taken simply to mean any graph with bars, most middle school curricula use “bar graph” to denote a graph with bars, a nominal X axis, and a quantitative Y axis; other representations which also use bars are given different names (for example, “histograms”)

Footnote 2: An error similar to this one was documented in Lehrer and Schauble (2001); in that case, students transferred one nominal characteristic into their representation of an ordinal variable.