“It’s Harder than We Thought It Would be”: A Comparative Case Study of Expert–Novice Experimentation Strategies

CINDY E. HMELO-SILVER, ANANDI NAGARAJAN
Graduate School of Education, Rutgers University, New Brunswick, NJ 08901-1183, USA

ROGER S. DAY
Pittsburgh Cancer Institute, University of Pittsburgh Medical Center, Pittsburgh, PA, USA

Received 1 January 2000; revised 8 December 2000; accepted 15 March 2001

ABSTRACT: Scientific inquiry is a complex skill. Aspiring physicians need to learn these skills so that they can be educated consumers of medical research as well as being collaborators in different kinds of clinical trials. But school science often fails to provide the kind of authentic tasks needed to help students develop appropriate reasoning skills and epistemological beliefs. In this study, we compared a group of expert cancer researchers with four groups of fourth year medical students (the “novice” groups) engaged in the task of designing a clinical trial to test a new cancer drug using a computer-based modeling tool, the Oncology Thinking Cap. Although the experts and novices reached similar endpoints, their reasoning processes differed considerably. For the experts, this was a task that required learning about the drug they were testing. The novices needed to learn about designing clinical trials, particularly about how variables interacted with each other, as well as learning about the drug. One of the major lessons learned by the novice student groups was just how complex clinical trial design really is. © 2002 Wiley Periodicals, Inc. Sci Ed 86:219–243, 2002; DOI 10.1002/sce.10002

INTRODUCTION

Scientific inquiry is a complex skill. Aspiring physicians need to learn these skills so that they can be educated consumers of medical research as well as being collaborators in different kinds of clinical trials. But, as Chinn and Malhotra (2000) note, school science often fails to provide the kind of authentic tasks needed to help students develop appropriate reasoning skills and epistemological beliefs.

Contemporary work in science education is framed by a view of science as inquiry and that science curricula should reflect what scientists actually do (Gitomer & Duschl, 1995).

An earlier version of this paper was presented at the 1999 Annual Meeting of the American Educational Research Association Montreal, Canada.

Correspondence to: Cindy E. Hmelo; e-mail: chmelo@rci.rutgers.edu
Contract grant sponsor: National Cancer Institute.
Contract grant number: 1R25CA63548-01A2.

© 2002 Wiley Periodicals, Inc.
This implies that learners should construct knowledge through investigation procedures that are similar to those that scientists use (Abrams, 1997; Gitomer & Duschl, 1995). There is mounting evidence that engagement in inquiry is motivating and can lead learners to a deeper understanding of science content and process (Krajcik et al., 2000; Minstrell, 2000). Learners need to gain experience in participating in scientific discourse, and using the specialized procedures of inquiry. By incorporating collaborative activities into inquiry, students must explain their understandings, argue with evidence, and critically evaluate the scientific explanations of others. Inquiry promotes development and transformation of ideas as well as helping them understand how scientific knowledge is generated as learners try to make sense of their observations. However, the demands placed on learners by inquiry required considerable support to allow them to both do and learn from these complex activities (Hmelo & Guzdial, 1996; Krajcik et al., 2000). The inquiry environment studied in this research provides such supports based on a cognitive science perspective on expertise.

Research on expert–novice differences consistently demonstrates that expert knowledge and strategies are more flexible than that of novices (Feltovich, Spiro, & Coulson, 1997). The scientific laboratory is no exception (Dunbar, 1993). In this study, we compared a group of expert cancer researchers with four groups of fourth year medical students (the “novice” groups) engaged in the task of designing a clinical trial to test a new cancer drug. We examined their knowledge and strategy differences as participants designed clinical trials, ran their trials on a simulator, and iteratively redesigned their experiments until they were satisfied with their results. This study adds to prior work on expert–novice differences because this is a more dynamic task than is used in typical studies of expertise. In addition, it is a task that is challenging for both the experts and novices, despite being conducted in an environment designed to support students’ learning about clinical trial design. This research focuses on the process of inquiry, as reflected in the experts’ and novices’ collaborative discourse, as they design and interpret the results of a Phase 2 clinical trial to test a new anticancer drug.

The Clinical Trial Design Process

Although Baker and Dunbar (1996) have demonstrated the existence of experiment schemas, they also speculated that different kinds of experiment schemas exist for different kinds of domains. For example, in developing new drugs, there are many different types of experiments that are conducted. A drug must go through several stages of laboratory and clinical testing before it can be available for general use (Simon, 1993). Phase 1 of clinical testing involves a small number of patients and the aim is to identify a safe dose. A Phase 2 trial may be subsequently conducted to see if there is any clinical response to the drug. In the design process, the researchers choose a single dosage and schedule for the drug. The researchers specify several minimal operational characteristics of the study and a design is chosen to meet those criteria. Typically, the goal is to distinguish a low response rate from a high response rate with good accuracy. A Phase 3 trial compares a new drug against existing treatments to determine whether it is more effective. In this study, we focus on designing Phase 2 trials.

In a Phase 2 trial, there is no control condition (see Figure 1). Instead, historical information about patient response and response to other treatments is used to determine the statistical criteria for declaring a drug worthy of further study. In addition to these statistical considerations, one of the keys to designing an effective trial is realizing there are dual goals for the study: (1) obtaining positive patient effects and (2) minimizing negative side effects. This occurs through a subtle interplay between the dosing, schedule, and conditional rules that allow treatment to be systematically modified in response to toxic side effects. In this
study, participants used a computer program that allowed them to design and simulate a Phase 2 trial. This provided the opportunity to observe the strategies that were used by expert and novice groups in undertaking this task and designing an optimal trial.

A cognitive model of the Phase 2 clinical trial design process is shown in Figure 1. The clinical trial design task requires a complex set of inferences to interpret the results and to plan future experiments (Chinn & Malhotra, 2000). There are multiple causal paths that can cause patient responses that lead to tumor shrinkage. In addition, the variables interact with each other. The amount of drug, the frequency with which it is given, and any
resting periods between courses of the drug all affect how the drug works and the intensity of its side effects. For example, a high dose of the drug will cause many side effects so a rest between courses may be needed so that patients can recover from the side effects. Moreover, the experimenter should set contingency rules, which are variables that strongly affect the dose and schedule. As part of a trial there may be a rule that says, for Grade 3 Hematologic (i.e., blood) side effects, decrease the dose by 50%, thus individual patients may receive reduced dosages from what the trial designers originally anticipated. If the patients experience these toxic side effects, they may not actually receive the dosage that the trial designers anticipated. Although there are standards for what constitutes a response, the measurement of many of these variables is complex. For example, a complete response is defined as a 50% shrinkage in a tumor that may need to be measured by a CT scan or other X-ray. The effects of the treatment are not perceptually obvious and must be measured indirectly (Chinn & Malhotra, 2000). In addition, designing a clinical trial also requires considering what other indicators to use to monitor patient well-being (e.g., survival time, blood counts, and quality of life indicators). Moreover, these variables must be measured at multiple intervals since times to response and toxicity vary with individuals. An environment that supports learning how to design clinical trials needs to include this complexity in a way that makes it manageable by novices.

**Expert – Novice Differences**

A large body of research has demonstrated that experts and novices differ not only quantitatively, in terms of how much they know but also that they differ qualitatively (e.g., Chase & Simon, 1973; Chi, Feltovich, & Glaser, 1981; Feltovich & Barrows, 1984; Glaser & Chi, 1988; Lesgold et al., 1988; Patel & Groen, 1986). Glaser and Chi (1988) summarize these differences as follows:

- Experts excel largely in their own domain because they have a rich base of domain knowledge.
- Experts perceive meaningful patterns in their domain of expertise reflecting a well-organized knowledge base.
- Experts are fast and accurate at solving problems within their domain because with practice, many skills have become automated, freeing up cognitive resources for processing other aspects of the task.
- Experts represent problems at a deeper level than novices do because of their superior conceptual understanding.
- Experts spend a great deal of time analyzing and representing a problem before they start solving it. This provides the experts with a cognitive representation that allow them to infer the relevant relations and constraints.
- Experts have strong self-monitoring skills.

Experts are also more adaptable to changing contexts (Feltovich, Spiro, & Coulsen, 1997). When this occurs, it can be attributed to expert memory for highly differentiated schemas or alternatively to a greater capacity to engage in reasoning from fundamental principles when schema-driven reasoning fails. These two types of expert reasoning “enable a rich knowledge-based flexibility” provided the expert is reasoning within his or her usual, stable domain (p. 133). For this latter type of reasoning to occur, experts must be aware of when they have to switch from their usual schemas to first principles (and this does not always occur). Feltovich, Spiro, and Coulsen (1997) argue that for this kind of flexible reasoning to occur, people need to experience the complexity of real-world knowledge application and
to understand the relationships between abstract knowledge and the messy world in which it will be applied. This understanding and flexibility may be uncovered by comparing expert and novice performance on a dynamic task such as designing experiments and interpreting the results.

Expert scientists come to an experimental design task with a well-defined schema for typical types of experiments (Baker & Dunbar, 1996). A schema is a knowledge structure that specifies the general properties of an object or activity but leaves out the specific details. These are the slots that are filled for a particular occurrence of an object or activity. Experiment schemas include slots for experimental manipulations, the types of dependent variables that are appropriate to measure, and conditional rules for the particular type of experiment being conducted. They also apply effective strategies to an experimental design. They know that scientific experimentation is most likely to lead to new understanding when the experimenter uses a strategy of varying one thing at a time, a strategy we would expect to see among experts (e.g., Trickett, Trafton, & Raymond, 1998). These schemas are necessary to manage the cognitive load demanded by authentic experimental tasks.

As well, experts can use analogies to other experiments that they have run or to other domains they are familiar with as well as having a well-connected network of domain knowledge. This task of designing an experiment is much more difficult for novices as they still need to construct these cognitive resources.

Dunbar’s approach to studying expert scientific reasoning has been to study practicing scientists working on real scientific problems in a social context (Dunbar, 1993). This work is unique in focusing on authentic tasks. In this paper, we compare expert and novice scientists working on an experimental design and simulation task. We maintained the social context by having experts and novices work in groups to design a special kind of experiment, the Phase 2 clinical trial in a complex task environment. The Oncology Thinking Cap (OncoTCAP) clinical trial wizard used in this study (described in the materials section) maintains many of the features of authentic experimentation that are needed for novices to develop sophisticated scientific reasoning skills (Chinn & Malhotra, 2000). Students have to plan sophisticated treatment strategies that involve an interaction among drug dosage, frequency of administration, and recovery time from toxicity. Moreover, they need to develop contingency rules to deal with toxic side effects (and lack of clinical effects). Because variables interact with each other, learning environments for teaching clinical trial design need to help students learn the complex sets of inferences required to control variables. Because of the delicate balance between positive and negative outcomes in these trials, students cannot just look at an end result; they need to examine patient histories to see these effects and to understand when to discontinue treatments to avoid irreversible negative outcomes. The OncoTCAP clinical trial wizard provides the opportunity for students to experience many of these characteristics of real clinical trials while providing support for learning. In this study, we analyzed the clinical trial design process by examining the actions that experts and novices made and the discourse they engaged in as they used OncoTCAP. In particular, we were interested in understanding the differences in knowledge utilization, problem representation, scientific reasoning strategies, and metacognition.

METHODS

Participants

The participants in this study were one group of experts and six groups of novices. The experts were two experienced cancer researchers, one a laboratory scientist, and the other a

1Because of poor audio quality for two of the groups, only four of the audiotapes could be transcribed for the qualitative analyses presented.
clinical researcher. The novice groups consisted of 24 fourth year medical students divided into six groups of four students. The student groups had received a lecture on clinical trial design previously and had attempted to design this trial on paper prior to coming to the computer session.

Materials

The participants’ task was to design a Phase 2 clinical trial. They accomplished this using the OncoTCAP and the Phase 2 clinical trial wizard interface (Day et al., 1998; Hmelo et al., 1998). All participants received the results of both laboratory tests of the drug, Pittamycin, and the Phase 1 clinical testing. This information included the maximum dose that was tolerated by patients (MTD) and the types of toxicities that were observed.2

OncoTCAP is a computer-based modeling laboratory for conducting experiments in cancer biology. It models the heterogeneous populations of cells that comprise tumors. OncoTCAP can model the important concepts in cancer biology such as cell cycle control; cell growth, death, and repair mechanisms; mutational processes, treatment characteristics, resistance, and schedules; and genetic characteristics (Day et al., 1998). OncoTCAP models these processes by specifying the different properties of cancer cells such as their genetic make-up, location, and resistance (Ramakrishnan et al., 1998). To make this tool easier for novices to use and to help them learn how to design Phase 2 trials, we developed the Phase 2 clinical trial wizard (see Hmelo et al., 1998 for additional details).

The wizard helps divide the task into four steps that lead the user through the experimental design process (Figure 2). In these screens, the user can enter the various design parameters for the trial such as the Schedule, Dose Modifications due to toxicity, Off treatment criteria, and Statistical Criteria. After completing these steps, the participants run this simulation in the Multiple Patient Simulator (Figure 3). The multiple patient simulator displays summary statistics for the trial, individual patient histories, and links to alternative representations for visualizing both types of information. The users can then return to the wizard to make further modifications to the design and re-run the trial. The summary statistics for the trial include the number of patients who had complete or partial responses, the number of patients who died because of either tumor or toxicity, as well as those who were still alive either with tumors present or no evidence of disease. The individual patient histories provided a chronology of events such as treatment and toxicity, the timing of drug administration, associated cell counts, and tumor location. Users could also see this information presented in graphical form.

Procedure

Participants were told that they were meeting to design a Phase 2 clinical trial for a new drug called Pittamycin. They were given the preclinical information and the results of a Phase 1 trial. The students initially met in unfacilitated groups to design the trial on paper. Participants used the wizard to revise and simulate the Phase 2 trial. The first author facilitated the group’s work by (1) helping participants with any interface problems they might have had and making sure that they understood relevant software features, (2) asking them to justify their changes, and (3) encouraging them to reflect on what they learned from this experience.

2 The two types of toxicity that were observed during the Phase 1 trial were neurological and hematological toxicities. Neurological refers to the effects on the central nervous system and brain. Hematological refers to effects on the developing blood cells and can lead to a decrease in red cells (anemia), platelets (causing impaired blood clotting), or white cells (causing an impaired immune system). All toxic side effects are graded in severity from 1 being mild to 5 being fatal, Four is considered life threatening.
experience. All sessions (both for the experts and students) lasted for approximately 2 hours. Sessions were audio- and videotaped and subsequently transcribed. In addition, OncoTCAP printed out each experimental design and the students printed out the final results.

**Coding and Analysis**

For all groups, the number of trials run and the number of variables changed per trial were counted. The number of variables changed was calculated in two ways. First, the absolute number of items changed was counted. Second, the absolute number was reduced by removing (a) changes that returned to a previous state in the experiment space, and (b) changes that maintained the same dose by varying the frequency (in other words if the dose given was increased and the frequency decreased resulting in the same amount of drug within each course of treatment, only one change would be counted rather than two). Because there was only one expert group, only descriptive statistics are presented here.

For the expert group and four novice groups, the transcripts of the session were analyzed using coding categories designed to compare and contrast how the groups went about designing their experiments and interpreting their results. This was based on the literature on expertise and scientific reasoning. Each turn was coded for each of the categories described below.
Multiple Patient Simulator summarizes the results of the simulation and allows users to examine individual patient event histories.

**Problem Representation and Goal Structure.** Because experts spend a good deal of time constructing an appropriate problem representation, we compared how the experts and novices defined the problem and the goals that they were addressing. In particular, we looked for evidence that the participants considered one or both goals of a Phase 2 trial: maximizing clinical responses (i.e., improving patient status) and minimizing drug-related toxicity (i.e., dangerous side effects).

**Use of Prior Knowledge and Analogies.** Dunbar and others have noted that experts make extended use of their prior conceptual knowledge and experience and that analogies are often productive (Dunbar, 1993; Nersessian, 1995). In Dunbar’s analysis of scientific reasoning, analogies were an important source of knowledge and conceptual change. We coded the protocols for the types of knowledge that were used and the use of different types of analogies. Dunbar (1993) identified two kinds of analogies that are used in planning experiments: regional and local analogies. Local analogies are comparisons made to similar situations within the current task-domain whereas regional analogies are comparisons made to domains other than the current task. Local analogies are commonly used to solve problems with experiments (Dunbar, 1993). Regional analogies occur much less commonly and involve mapping over a system of relations when researchers need to fill in gaps in their own knowledge and plan new experiments. In this study, local analogies refer to previous trials within the computer session whereas regional analogies refer to other studies or drugs. In addition, we coded the use of other kinds of prior knowledge. Participants were coded as using prior subject matter knowledge if they referred to conceptual information that...
they learned from books or in their classes prior to coming to the computer session. They were coded as using *prior experience* if they referred to actual experiences doing the same or similar tasks. These were coded as prior experience rather than analogies because they referred to very general statements such as “We’ve seen that before” or “This is typical.”

**Metacognitive Control.** Experts are better than novices at monitoring and evaluating their performance as well as reflecting on their actions (Ertmer & Newby, 1996). Reflection helps people learn which strategies are more or less effective in designing trials and in understanding the nature of the drug being tested. Metacognitive statements were coded in three categories: monitoring, evaluating, and reflecting. Because planning reflected not only their metacognitive control but also their scientific reasoning, we include planning in the scientific reasoning category. Statements were coded as *monitoring* if they indicated an assessment of their ongoing progress. This included an awareness of making an error or that someone else in the group was making an error as well as an understanding of their current state of knowledge. **Evaluation** refers to evidence that the group was making a value judgment by relating an outcome to the actions that led to that outcome. A turn was coded as *reflection* if it included a discussion of whether actions were leading to their desired goals and the effects of earlier actions. Reflection occurred as participants considered the earlier plans they made and strategies they used, their evaluation of the effectiveness of those plans and strategies, and the resulting revisions to their thinking.

**Scientific Reasoning.** This includes many phases of the scientific inquiry processes such as how groups went about planning their experiments, generating and rejecting hypotheses, making predictions, interpreting the results, and making decisions about actions to take. In addition, it includes controlling variables and deciding when they were finished with the task. Experienced scientists should be quite deliberate in planning their experiments once they have represented the problem that they are trying to deal with (Lesgold et al., 1988). We looked at three planning strategies: theory-driven planning, data-driven planning, and unjustified planning. Ideally, scientists should plan their experiments with appropriate theories in mind and these theories should drive their plans for subsequent actions. It is also possible that novices, without a global understanding of their task, may plan based on the most recent available data, in other words, their plans may be *data-driven*. The results of data-driven planning are very specific plans to change a variable based on evaluating previous results. Turns were coded as *unjustified planning* when no explanation for the plan was described. This was more of a trial and error approach that was not rationally grounded.

Good scientific reasoning requires that one design experiments to test specific hypotheses derived from theories. But as Klahr and Dunbar (1988) have shown, often experimentation proceeds independently of hypotheses. To examine this we examined the transcripts for evidence of specific predictions and how results were interpreted.

With effective reasoning, participants should generate predictions and interpret their results in light of those predictions. These were coded as *high-level interpretations*. Often, participants just discussed low-level interpretations in which they interpreted the software results without considering the broader implications of those results. *Low-level interpretations* were very literal summaries of displays such as “There were a lot of toxicity deaths here.”

A hallmark of expert experimentation is careful control of variables by changing only one variable at a time (Trickett, Trafton, & Raymond, 1998). This is critical because if more than one variable is manipulated between trials, it is difficult to make causal conclusion about the results. Given the scientific background of the fourth-year medical students, we would expect both experts and novices would only change one variable at a time between trials, but it
was also possible that the experts and novices might have a different understanding of what actually constitutes a variable in this context. Thus, we examined both the transcripts and the software logs for indications of variable control. Finally, the transcripts were coded to see the extent to which the participants settled for a “good enough trial,” that is whether they satisfied or whether they tried to optimize their research design. The issue of satisfying versus optimizing is important as it gives an indication of the group’s overall goals, motivation, and cognitive burden. If the group is driven by a task-completion goal versus a learning goal, there should be differences in task-persistence. A good scientist explores many possibilities before settling on a particular design (Chinn & Malhotra, 2000). However, even after considering possible constraints of time and task demand, we were interested in identifying differences between experts and novices.

Reliability. To check the reliability of this coding scheme, two independent raters coded the complete transcript of one group. Inter-rater agreement was calculated in two ways. First, we calculated agreement by comparing only those turns that were actually coded. Second, we calculated agreement on all turns. The first, conservative method indicated 78% agreement. The more liberal method indicated 99% agreement. The more liberal method indicated 99% agreement.

RESULTS AND DISCUSSION

The expert and novice groups generally reached similar endpoints in their trial design. The modal endpoint for both was a fairly high dose of the drug with some resting time that allowed patients time for the toxic side effects to resolve. The process by which the experts and novices got to that endpoint differed considerably, however. The expert group ran many more experiments than the novices did (17 vs. a mean of 8). As well, the experts were systematic, changing a mean of 1.94 variables between trials. However, when dose-preserving changes were factored out, this mean dropped to 1.06 changes/trial. A dose-preserving change was a return to an earlier state (e.g., the same dose and treatment schedule as had been used three trials earlier) or those that maintained the same total dose. These latter occurred, for example, when the experimenters increased the dose from 60 to 120 but reduced the frequency from 4 to 2; in both cases, the patients received 240 mg of the drug within a course. The students talked about systematically controlling variables, but in fact they changed a mean of 3.18 variables in each experiment. Even considering returns to earlier states and dose-preserving changes, the novices still changed 2.83 variables between each trial.

The types of changes that the experts and novices made differed as well. Although there was little difference in the absolute number of changes made (22 for the experts vs. a mean of 19 for the novices), the locus of these changes differed. Both the experts and the novices made changes in the area of dose modifications in response to drug toxicity (35 vs. 28%). The novice groups made another 28% of their changes in specifying the circumstances under which patients would be removed from the trial. The experts included a dose modification rule from the beginning whereas none of the student groups included this until the second or third trial design. The expert group only made 3% of their changes in this area. The experts spent the majority of their effort in adjusting the dosage and drug schedule (61%), and although this was an important focus for the novices, they only made 45% of their changes in this area. The experts were more tolerant of toxicity than the novices. In their final design, there were 12 patient responses and seven deaths because of toxicity. Once students understood the minimize toxicity goal, they became more risk averse. Their final designs had a mean of 7.5 patient responses and 2.2 toxicity deaths. To help in understanding these quantitative indicators of performance, the qualitative coding scheme was applied to the data to generate frequency counts to closely examine the experimental design and interpretation.
process. To help illuminate this analysis, we present excerpts from the expert and novice (student) transcripts.

The experts generated a total of 812 coded turns whereas the student groups generated a mean of 2,143 coded turns. There were many more turns in the student group because there were four students per group compared with two in the expert group. Because of the differences in the number of turns, all results are reported as percent of total turns or as percent of a specific type of code within turns in a category (e.g., references to local analogies within the category of prior knowledge).

For both the expert groups and novices, the largest number of turns was in the areas of metacognition and scientific reasoning (Figure 4). The students spent more time engaging in metacognitive activity, though as we discuss later, the type of metacognitive activity differed between the groups. The experts were more likely than the novices to make references to prior knowledge (including the use of analogies), to discuss the goal structure of the task, and to attend to scientific reasoning. In the remainder of the results section, we discuss the nature of the discourse within each category and provide examples from the expert group and from the student groups.

**Representing the Problem: Task Goals and Schema**

The expert group spent considerably more time than student groups in discussing how the structure of the problem was represented in the dual goals of the task (Figure 4). Although the number of turns spent discussing the goals was greater for the experts, the distribution of those statements did not differ, as shown in Figure 5. The experts and students both spent approximately 33.5% of their goal discussions focused on maximizing responses and 66% on minimizing toxicity. For the experts, however, considering the minimizing toxicity goal came earlier in their experimentation.

For the experts, this was a familiar problem-solving task. They understood the Phase 2 trial design process so the only unfamiliar part of the task was understanding how the drug, Pittamycin worked. For the novices, both the task and the drug were unfamiliar. This was
reflected in how the two groups represented the problem and the underlying goal structure. For the experts, the slots in the wizard represented familiar variables that had to be considered in designing a clinical trial. The novice group, with their limited prior knowledge, engaged more in understanding the variables and spent much more time in constructing a shared understanding of what the variables meant (such as what defines a complete response). The contingency rules were a frequent item for discussion and action although the novices did not initially understand that these were variables and that these interacted with other variables.

Despite having attempted this task on paper, the novices still found this task novel and complicated as they tried to determine how a response was defined as J said:

> Ah and we think this... we think fifty would... something in that range would give us a good chance cause you know... we're not trained in that area. I'm not sure if we have enough, ah, I'm not sure if we have enough independent variables picked out in order to make that kind of determination.

Here, J has tried to decide what the criterion for a complete response (CR) to treatment would be and perhaps a 50% reduction in tumor size would be appropriate. This is the standard criterion that is used in a clinical trial, but then J gets confused and decides that this is somehow related to the number of independent variables.

One of the keys to representing this task appropriately is to understand that there are two goals: patient improvement (defined as complete or partial response to the drug) and minimizing side effects or reducing the occurrence of toxicity. The experts clearly understood this from the start:

> A: Cause I guess the issues here are toxicity issues, not really response issues.

> C: They're both.

> A: Well, yeah they are kinda both.
Though it appears that expert A was unclear about the dual nature of goals, all of their actions represented a cycling between these goals as they went between adjusting the drug dosage and treatment schedules and making changes to the dose modification rules. Trial designs were debugged with the two goals in mind. For instance, if they found that, in a trial, there were several toxicity deaths because of neurological or hematological toxicity, their actions were geared to reducing the toxicity levels. At the same time, they maintained the goal of obtaining a good response rate (“CR” refers to a complete response). At the very end, as they analyzed their design the experts noted:

A: This is an impressive drug. I’m excited by this drug, everyone’s recurring but I am excited about it.

C: It’s alright.

A: CR rate is 50%.

C: Toxicity is low.

A: Yeah alright!! A winner.

Thus, they assessed the effectiveness of the trial based on the dual goals of maximizing responses and minimizing toxicity.

The novices used the objectives from their written instructions to get started with the task. They clearly indicated their focus on the single goal to obtain a clinical response.

Cl: We decided that . . . based on the objectives of the . . . this task . . . we are going to measure it like two things in terms of ah . . . measuring the clinical response of Pittamycin to breast cancer. One being ah . . . we’re going to use imaging studies to . . . to monitor that and two being ah, serum levels of possible ah, breast cancer ah, markers.

Although they allude to patient monitoring, it was for the purpose of assessing response to treatment rather than looking for toxicity. They continued to focus on the single goal for a while:

J: Ah yeah . . . what will make a successful trial uhm . . . Well in a individual patient, we wanted a 20% ah . . . to ah . . . to define a clinical response.

Students did not make any dose modification rules related to toxicity in their first two trials. It was only after they analyzed the effects of the drug on individual patients, that they began to construct knowledge of the toxicity effects; it was only at this point that they began to consider the second goal:

J: We could, we could lose people to toxicities. We could lower the dose and, and increase the ah . . . We could either lower the dose or increase the length of the trial . . .

In this last segment, following the second simulation run, the students began to think about the importance of taking action to minimize toxicity. In the segment that follows, they began to formulate a plan for action and discussed the tradeoff between responses and toxicity:

S: Yeah. I think we need to play with the toxicities. Reduce the people . . .

J: We could oh reduce the dose while we get the toxicities.

S: Yeah . . .
J: Yeah, we can do that.

S: Things like that . . . cause we keep it in neurologic hematologic toxicity and people aren’t completing the trial. We still got some responders though even though we were taking half of everybody. And some of the people had high toxicity. But the toxicity doesn’t seem to be killing people.

S and J were having a discussion about needing to modify the doses in response to toxicity. They observed that about half the patients were being taken off the trial because of high toxicity and so were not getting much treatment. This indicates a developing concern for reducing the toxic side effects in addition to the already articulated goal of obtaining patient responses.

It is the experts’ rich prior knowledge and experiential base that accounts for their different understanding of the task. Before starting the computer session, the novices had mostly abstract knowledge of the Phase 2 clinical trial design process whereas the experts had experiential knowledge from designing many previous trials.

The Role of Knowledge and Strategies

In examining the transcripts, we explored the role of prior knowledge, specifically, the use of local and regional analogies as well as the explicit use of abstract knowledge as well as general references to prior experience. In general, the experts were slightly more likely to refer to prior knowledge than the novice students were (3.69 and 1.57%). The content of the discourse was considerably different for the experts compared with the novices (Figure 6). The experts’ use of knowledge was largely experience-based as shown by their references to prior experience (40%), local (10%), and regional analogies (33%). In contrast, the novices resort to prior conceptual knowledge and facts (44%), local analogies (31%), regional analogies (21%), and only rarely prior experience (4%). For the novices, the only

![Figure 6. Prior knowledge use.](image-url)
experience they had to use was from other trials they ran during this computer session, accounted for by the relatively large proportion of local analogies.

Their prior experience provided the experts with well-established schema-based strategies to focus on the task. The experts had a schema for a typical trial that they used to identify whether their trials were consistent or inconsistent with that schema. For example, while deciding on off-treatment criteria expert A notes that “25% is standard for progression.”

He was also able to compare this to other trials he has participated in when he says “so it’s not like a typical Phase 2 basically...” As the software computed the number of subjects needed for their trial design, one of the experts noted “You put the Simon Formula in there?” noting that the computation was consistent with what was normally obtained by hand calculation.

The experts’ prior experience was also reflected in their use of both regional and local analogies. The experts used local analogies as they compared results from a later trial with those of an earlier trial: “So the first time we did it, it was good (response) but the toxicity was bad.” These seemed to be most useful in assessing how well they were doing in the experiment space. Regional analogies, generally between the drug being tested, Pittamycin and other drugs, seemed more productive in designing further trials. For example, one of the experts made an analogy to another drug (nalaphene) as they are trying to determine the appropriate dose modification: “I mean what do we do with nalaphene? We hold the dose.” The regional analogies were also used to help them interpret the results of trials “you know what this reminds me of? IL-2. You live and you get a CR or you die.” They made similar comparisons to another drug, cisplatinum, as they noted both the high response rate and toxicity and got excited about the effectiveness of the drug:

A: Gotta reduce the toxicity.
C: What sort of deaths, toxicity death rates are there?
A: I bet you the hem, the hem, well, this is too much.
A: Although look this is like 20%, this is 50% CR rate... it’s like, it’s like cisplatinum.
C: It’s a wonder drug.

Novices had limited prior knowledge and/or experience with this task, and therefore were unable to access information that could be useful in solving this problem. While defining a partial response, the novices got bogged down in understanding the conventions for determining a response:

J: For a partial response, we wanted at least a 20% decrease in bi-directional area as determined by ah... ah... unending studies.
Facilitator: Where did you come up with that number?
J: Previous... uhm.
Cl: I kind of think that 20% was arbitrary.
J: Arbitrary?
Cl: Yeah.

3Off-treatment criteria are used to decide when to remove the patient from the trial. Examples of these include disease progression and severe toxicity.
However, they did make reference to a web page that recommends the standards used for a starting dose as being a fraction of the maximally tolerated dose (MTD) that was identified in earlier studies, so they did bring some prior knowledge to bear on the task as CI noted “that should def... that NCI web page recommend 80% of the MTD.”

Despite a reliance on prior knowledge, they were quite aware of the limitations to their understanding, particularly concerning the variables involved and their relationships to treatments and toxicities.

J: You know, I didn’t expect to be, you know, looking at when they had their toxicities in relation to the treatments and... ah that was probably... you know... that one complicated variable and probably others that I didn’t expect to have to deal with... you know I was just picking.

Their limited knowledge in this area was reflected in their lack of regional analogies to other drugs. However, they did make local analogies within the task by comparing results from one trial to previous trials.

Metacognitive Activities

The experts and students did not differ much in the percentage of turns spent on metacognitive activity (12.44 vs. 13.67 % of total turns), but the nature of their metacognitive activity differed (Figure 7). Neither group spent a great deal time reflecting (12.87% of metacognitive activity for experts vs. 11.60% for students). What is striking is the difference between monitoring and evaluation between the experts and the students. The students devoted considerable effort to monitoring their cognitive activities (66% of all metacognition) but spent considerably less time evaluating their progress (22%). The experts were more likely to evaluate their progress (47.5%) rather than just monitor it (39.6%). One explanation for this may be that the students simply did not have the knowledge and experience to evaluate their

Figure 7. Metacognitive activity.
progress. An alternative explanation is that the cognitive demands of the task were high enough to interfere with the students’ capacity to evaluate their progress. Another possible explanation could be that since the students had to work together on an unfamiliar task, there was additional monitoring to make sure all group members were at the same level of understanding. For the experts, joint understanding was almost instantaneous since the task was a familiar one. Also, experts could evaluate the results by comparing it to other drugs or studies, students were limited to making comparisons across trials. It was only after working on the task for a while that the students developed an understanding of what could be evaluated as a good action and otherwise.

The experts frequently monitored, evaluated, and reflected on their performance. They kept a good mental account of the effects of their actions and revisited information from previous trials to make decisions about future trials as shown in the following examples interspersed throughout the transcript:

How many CRs did we have the last time?

Why don’t we keep the same parameters, go back to 60, cause 60 q month really was good.

So we know we originally got a big CR, but a lot of tox, so how can we adjust it?

So I mean, we like the weekly schedule, it’s clearly the best in terms of response.

The first sentence shows an example of the experts monitoring their performance whereas the second and third statements show the experts evaluating their activities. Their reflection, shown in the last example was used to drive their actions on subsequent trials. The experts’ prior experience had allowed them to internalize the external standards that are often used to evaluate trials. The next example shows that they are quite aware of those standards as they evaluate their results and demonstrate a shared understanding of what it means for the number of patients dying to be “too high”:

A: Death! They all died, oh man, CR though, hey, oh this is terrible, man, the FDA is after us big time, oh we’re in big trouble.

C: That’s way too high.

They are reflective as they try to draw lessons from earlier trials at many points during the computer session for example “Why don’t we say this? Why don’t we keep the same parameters, go back to 60, cause 60 q month really was good, I mean, that’s even if there was a good response, don’t you agree?”

These kinds of monitoring and evaluation statements were observed less frequently in the novice groups, probably because their cognitive resources were fully engaged in understanding the task and interpreting the results of their trial. The students’ monitoring helped raise their awareness of what they did not know, an important aspect of self-regulated learning, as S notes his limited understanding of the statistics involved:

So it says, yeah, alpha error at point zero five (.05) means a 5% chance of accepting vitamin D-3 is a good drug if the true response rate is 5%. Beta error, point two zero (.20) means 20% chance of rejecting it if the true response rate is 20% . . . I don’t understand what that means.

In this example, they were constructing self-explanations as they try to connect what they do know to the statistical information appearing on the screen (Chi et al., 1989). In
addition, they were identifying possible goals for future learning. The students later came to recognize what made a drug a “bad drug”:

J: That’s a bad drug.

S: We couldn’t have found it unless we did this.

Here the students acknowledge their limited understanding of the task, hence their limited capacity to evaluate results.

Although reflection was infrequent, for the students, it provided them with opportunities to consolidate what they had learned. Here one of the students demonstrated some insight about the effect of the conditional dose modification rules, noting, “But we only had, well, we had one person die from the neurologic toxicity. We had a lot of people not get all of the treatments because of hematologic toxicity.” These metacognitive discussions functioned as places where the students worked on constructing coherent understanding.

Scientific Reasoning

More of the expert dialogue focused on scientific reasoning (18.47%) compared with the novices (12.77%). The experts and novices focused on different aspects of scientific reasoning, shown in Figure 8. Although both groups engaged in planning during 22.5% of the turns coded in scientific reasoning categories, the experts devoted most of their effort to theory-driven planning (14.67%) and a lesser amount to data-driven planning (8.67%). The students rarely did any theory-driven planning (4.35%) compared with their data-driven planning (15.54%). Both groups did a large amount of low-level interpretation of the data screens but the experts engaged in nearly twice as much high level interpretation as the students (13.33 vs. 7.12%).

![Figure 8. Scientific reasoning.](image-url)
**Planning.** The experts were clearly theory-driven in their planning, primarily because their goal structure was defined from the beginning of the trial design process. Their actions were closely linked to both goals as expert A stated, “Right, I think they’re probably down to dying heme deaths, so we’ve got to adjust the heme tox.” The goal of minimizing toxicity was driving the change in the dose modification variable in response to hematologic toxicity. Moreover, they planned the next step in case the change that they made was not effective as they were trying to decide what toxicity level to use in invoking the dose modification rule:

Let’s do, what do you want to do? One? Or let’s do two? I don’t know. If we do one and it’s bad we can always go to two. OK, but if we have a low CR rate, we can go to 2.

When they found that an overly stringent dose modification rule also led to a decreased response rate, they cycled to the goal of maximizing response rate.

A: Alright, so now we got rid of toxicity, why don’t we change the dose modification again?

C: The patients are being dose-reduced.

A: Oh, they are?

C: Yeah.

A: So lets increase our, [dosage] but that’ll increase the toxic death rate too. That’s the problem.

Experts also did data-driven planning. This refers to planning specific task-related actions like setting dosage at 60 or toxicity level for dose modification at 3. This was more data-driven than what we described in the last paragraph. For example, in this excerpt, the experts observed a high number of deaths because of hematologic toxicity and realized the immediate need to modify their dose modification rule:

A: Still 50% toxic death rate.

C: OK lets see, what are they dying of? They have to die of hematologic toxicity.

A: Heme 5, all right, so we need to adjust the heme now.

This constituted a good deal of their planning so the experts used a combination of top-down goal-oriented planning and bottom-up data-driven planning. Regardless of the type of planning, the experts had hypotheses in mind as they made predictions about the effects of their changes:

S: Most of the deaths early are due to toxicity.

A: Interesting point, ok, that’s a good point, so if we reduce the tox, we’ll increase our mean survival.

Novices began the task without clear goals and discovered that there were toxicity issues to consider as well. Their planning and processing strategy was opportunistic in that decisions were made based on the latest information presented by the simulation. Once the novices understood the toxicity variable, they seemed to analyze further trials in the same way. They used the individual patient histories to understand the task and construct knowledge along the way. This elaborated analysis of the patients was rarely seen in the experts.
The novices tended to use a trial and error method when they were trying to figure out what would happen if they set certain parameters in a particular way. For example, here the novices were trying to decide what Pittamycin dose to use for their next trial.

J: Increase this dose. There it is.
S: Won’t we? Just whatever you guys want.
J: Yeah.
Cl: 25.
J: 50.
Cl: 50? No, no, no.
J: I hate . . . 42.5. God, you’re killing me.
P: It’s 25.
S: Put 25 in.
Cl: 25. 25.
J: Don’t . . .
P: Do you just want to make a decision here?
S: We’re on one.
P: We’re on one? See what happened.
J: For the next treatment?

In this example, they were throwing out a lot of numbers in an unprincipled manner to see what happened. The novices were more exploratory whereas experts seemed to know the purpose behind their action and made predictions about expected results. This unprincipled approach let to suboptimal experimentation strategies.

**Experimentation Strategies.** For experiments to be interpretable, the experimenter must change only one variable in each trial. That required an understanding of what constitutes a variable in this task domain. As noted earlier, the experts did understand this principle of experimental design quite well. Neither group made many statements in this area (experts: 0.67% of scientific reasoning statements, novices: 2.87%). When these discussions occurred between the experts, however, the content was focused and was reflected in their actions as they considered making several changes here:

C: Do you want, maybe we should treat more often.
A: Wanna treat 30? . . . Why don’t we do 30? Twice a week?
C: Yeah, let’s do 30.
A: Then change the hem parameter back.
C: So you want to do it here?
A: Yeah.
They started out considering changing the dosage and frequency (to keep the total dose the same), and then considered a change in the dose modification rule (the “heme parameter”). At that point, they were aware of changing too many variables and they left the dose modification rule unchanged.

In contrast, the novices considered variable control issues differently. While group members made mention of the need to change only one variable at a time, there was a mismatch between talk and action. This particular group changed an average of 2.83 variables between each trial (Range: 1–6 changes). This could well be due to the collaborative nature of the task or more likely, to how the students constructed what counted as a variable. In the example shown below, only one change was made for the second simulation. However, discussion about toxicity criteria started before designing the 3rd trial. It was at this time that a conversation on variable control ensued. In spite of the conversation, five variables were changed at this trial.

J: We can keep the dose the same or we can change the dose at each. We can change the dose. We can change the interval. Ah I don’t think we need? Ah do you wanna change . . .? Do you wanna make the life of this trial pretty long also, since we had a ah . . .

Cl: The eight-week.

S: That’s a lot of doses.

Cl: If it’s every 4 days . . .

J: But you know . . . or ? how? What do we wanna change? I mean what’s bothering us the most?

S: Well we made the study longer once now let’s make the study shorter and more . . . drug this time and the next time we could do both. Do you know what I mean? Let’s do the . . . Let’s do one thing . . .

Cl: One thing at a time.

S: One thing at a time.

J: Let’s do it all.

S: We can save toxicity changes to until next time . . . Let’s make one change give it like . . . You can give it a higher dose and more frequently.

J: There’re a lot of things we can change.

Further discussion about controlling variables followed:

J: If the toxicity is greater than or equal to three then we take them off so we know. So

S: I thought we were going to change that too? Well though we’re changing a whole bunch of stuff.

S: No we’re not. We’re changing one thing. It has two parts.

From this discussion, it is not clear if they realized that changing the toxicity via the dose modification rules was a variable that affected drug dosage. So although these novices have conceptual knowledge of the principles of variable control, they were struggling with
applying it to the experimental design task and understanding what counted as a variable in this situation.

Another issue in designing clinical trials is deciding on a good trial design. The experts continued to systematically experiment until they optimized their results. The experts discussed optimization infrequently but more than the students (3.33 vs. 1.38%). They exhibited indications of improving their design throughout the session but particularly so at the end. Even after attaining the goals they set, they acknowledged that their design could be improved:

A: This is an impressive drug, I’m excited by this drug, everyone’s recurring but I am excited about it.

C: It’s alright.

A: CR rate is 50%.

C: Toxicity is low.

A: Yeah alright!! A winner. Although it’s still pretty toxic.

A: Gotta reduce the toxicity.

Even after getting a good result, they continued to refine their trial even further to reduce the drug toxicity. They used an analogy to an existing drug, IL-2, as a standard to determine when their design was “good enough.”

A: So wait, now the tox is gone down, so should we try to refine it or are we happy with this?

C: This is better than IL-2.

A: I think we are satisfied with this. We appreciate the toxic death rate being better. And I think by refining the dose reduction, for that course, is probably how we could make it better.

Because the novices took a longer time to get to a satisfactory trial design and because they had less knowledge of how clinical trials normally worked, they devoted less attention to optimizing their design. After the 12th trial, they reached their best solution with C saying “Look at this. This is the best we’ve had so far.” However, they did continue to debug that trial and designed two more trials. Their last trial was worse than earlier attempts, but they did not improve it any further because of lack of time.

Predictions and Interpretation. Making predictions and working through the discrepancies between their predictions and their interpretation of the experimental results should help both experts and novices learn (Krajcik et al., 2000; Minstrell, 2000). For the experts, this learning was about the drug whereas for the novices, this learning was about the clinical trial design process. Predictions were discussed infrequently (experts: 4.00%, novices: 4.44%). The expert predictions tended to refer to specific types of patient outcomes as they predict the major cause of death to be heme toxicity, which they then turn into a plan for action, when their prediction is confirmed.

C: Oh, we’re killing people left and right.

A: Still 50% toxic death rate.

C: Ok, let’s see. What are they dying of? They have to die of hematologic toxicity.

A: Heme 5, all right, so we need to adjust the heme now.
The students’ predictions were often vague as S made a prediction about response to treatment “I like this combination maybe we’ll try this one again. I bet you it comes out better.” So here, the student was making a very general statement that this trial should be better than earlier trials without making any reference to what criteria were being used to define “better.”

Both experts’ and students’ interpretations were often low-level interpretation of the data (44 and 49% respectively) and for both groups, this was the focus of a substantial amount of their discussion. Where the two groups differed was in high-level interpretation, with the experts devoting a good deal more discussion to that than the novices (13.33 vs. 7.12%). The experts’ low-level interpretations appeared when the simulation results were coming across the computer screen as they made comments such as “Oh this is terrible, oh death, death look how many CRs though” and “So, the neuro tox killed them . . .” For the students these were literal interpretations with few, if any inferences as S said: “Well that’s what we were hoping for, yeah, and so far nobody had a greater toxicity at Grade 2. Lets keep going and see what kind of toxicity’s we can find in other patients. So this, that person had a neurologic injury, see here. Neurologic 1 is when they got toxicity and neurologic 0 is when it resolved. This is them over time and months. The tumors start growing at 0 months so that at hundred nine point seven eight (99.78) months is when they die.” Because of their unfamiliarity with the task, they devoted a good deal of their cognitive resources to learning how to interpret the display, and perhaps trying to learn which information was really important for this task. They were trying to relate this literal interpretation back to their predictions (though not necessarily to the actions they took).

The experts tried to make high-level interpretations to reach a general understanding about how Pittamycin could be used safely. After skimming a few patient histories, they came to a conclusion about reduced toxicity effects as C noted “See every other week is good for the toxicity obviously.” Here, again they made conclusions about the strategy they used “The intense therapy is what works, the problem is that it’s toxic.”

The students tried to make some of these higher level interpretations but they were not as adept as the experts were. Here, J was trying to make a high-level interpretation but he was not very confident:

J: Maybe he just died. For the first, I don’t know; this is just the first two patients. Maybe we didn’t . . . Maybe the trial wasn’t long enough, or, maybe. Even if the trial isn’t long enough for this drug, it won’t do any good, no matter how long we do it, but so far the toxicity’s are . . .

It was only after several experiments that the novices confidently made statements about how the drug works:

F: Now why is it that the ras mutated cells are more sensitive to Pittamycin?

J: That’s because the drug works by that mechanism.

Eventually the students constructed a better understanding of the drug and of the drug’s behavior as J notes ”It seems like we got all our production in the first four weeks.” Note that this is similar to the expert comment about intense therapy being more effective, but this high level interpretation process was much more effortful for the students.

CONCLUSION

The expert group and the student groups got to similar endpoints in their trial designs. They both ended up with large initial doses and at least some resting time between trials to allow the patients to recover from toxicity. All groups had rules to modify the dosage when toxicities were encountered though the novices were more stringent in invoking their rules.
But the task was quite different for both as was the process by which experts and novices arrived at their final design. For the experts, this was a dynamic problem-solving task in which they needed to learn about the characteristics of a new drug, and they used analogies to familiar drugs to help move their problem solving forward. They flexibly adapted their strategies and planning to the dual goal structure of the task. They reflected on how well their plans and strategies had worked. The novices did not have an extensive knowledge base and so they had to learn about the Phase 2 trial design process as well as about the characteristics of the specific drug. They did not use analogies very often and they did not have many cognitive resources available for being reflective. They did engage in the task with great deal of interest and persistence, and as we have demonstrated elsewhere, they learned a great deal about Phase 2 trial design (Hmelo et al., 1998). Their initial schema for an experiment consisted of a traditional experiment with a single manipulation. The complex simulation environment provided opportunities for students to apply what they had learned, and iteratively redesign based on feedback from the simulation. These results demonstrate the importance of having an authentically complex task environment as many nonsalient aspects of the experimental design process were highlighted. Moreover, the complexity forced the students to monitor and occasionally reflect on their performance. Part of what they needed to learn for this task was what counted as a variable and how variables interacted with each other, as one of the students commented during postsession reflection:

You know I didn’t expect to be you know looking at when they had their toxicities in relation to the treatments and . . . Ah, that was probably . . . you know . . . that one complicated variable and probably others that I didn’t expect to have to deal with . . . you know I was just picking . . . you know . . . Did they have a reduction of tumor and that’s it? And there are a lot of other . . . a lot of other things that we can use to vary, vary the performance.

As many of the students noted, this process was much harder than they thought it would be.

Scientific inquiry does not happen without a context in which to reason (Minstrell, 2000). As both Minstrell (2000) and Chinn and Malhotra (2000) note, experience with authentic materials, tasks, and phenomena provide contexts for scientific reasoning. As we observed in this study, the collaborative dialogue supported learning among both the novices and experts, helping them clarify their thinking and consolidate their ideas. Providing cognitive guidance in the form of scaffolding helped support student learning as it supported them in reaching endpoints similar to those of the experts.

We are extremely grateful to the experts and novices who participated in this study. We thank Sailesh Ramakrishnan, Bill Shirey, Adam Brufsky, Candace Johnson, Quinsou Huang, and Joseph Baar for their assistance with this work. We thank Cynthia Booke for doing the reliability coding.

REFERENCES


