A Collaborative Convergence on Studying Reasoning Processes:

A Case Study in Statistics

Marsha Lovett

Department of Psychology and Center for Innovation in Learning

Carnegie Mellon University

Running Head: Studying Statistical Reasoning

To appear in D. Klahr & S. Carver (Eds.) *Cognition and Instruction: 25 Years of Progress.* Mahwah, NJ: Erlbaum.

This chapter begins with a memory experiment, and you the readers are the participants! Please read the brief story below and try to memorize it. There will be recall questions asked later. (You may recognize the story, as it is quoted verbatim from an already published work [Gick & Holyoak, 1983].)

The General

A small country was ruled from a strong fortress by a dictator. The fortress was situated in the middle of the country, surrounded by farms and villages. Many roads led to the fortress through the countryside. A rebel general vowed to capture the fortress. The general knew that an attack by his entire army would capture the fortress. He gathered his army at the head of one of the roads, ready to launch a full-scale direct attack. However, the general then learned that the dictator had planted mines on each of the roads. The mines were set so that small bodies of men could pass over them safely, since the dictator needed to move his troops and workers to and from the fortress. However, any large force would detonate the mines. Not only would this blow up the road, but it would also destroy many neighboring villages. It therefore seemed impossible to capture the fortress.

However, the general devised a simple plan. He divided his army into small groups and dispatched each group to the head of a

> different road. When all was ready he gave the signal and each group marched down a different road. Each group continued down its road to the fortress so that the entire army arrived together at the fortress at the same time. In this way, the general captured the fortress and overthrew the dictator (Gick & Holyoak, 1983, pp. 35-36).

Introduction

This chapter focuses on the problem of improving young adults' statistical reasoning skills , with a particular emphasis on transfer outside the original learning context. Effective transfer is critical here because statistical reasoning is applicable across a wide variety of domains and in daily life; statistical reasoning skill is of little value if it can only be applied in the statistics classroom. And yet, students have great difficulty learning statistical reasoning skills in a transferable way (e.g., Garfield & delMas, 1991; Pollatsek, Konold, Well, & Lima, 1984). These instruction-oriented studies document that many current approaches to teaching statistics—even modern reform-based pedagogy—leave significant room for improvement, but they provide relatively little guidance on how to proceed.

Learning the appropriate representations for knowledge (not overly specific, not overly general) is key for effective transfer. Understanding how knowledge representations are formed and changed during learning is one of the

foci of cognitive theory. This suggests that a successful route to improving students' transfer of statistical reasoning skill may rely heavily on integrating instructional and cognitive theory, while still maintaining a link to the realities of the classroom. Unfortunately, the fundamental tension between theoretical and applied methods over the last thirty years has led to the emergence of three distinct approaches, each emphasizing primarily a single perspective:

(1) Develop cognitive theories to describe knowledge representations that explain observed performance on simplified statistical reasoning problems;

(2) Conduct empirical work to study students solving simplified statistical reasoning problems in real-world contexts.

(3) Work in the classroom, with all the complexity that doing so implies, to develop new instructional techniques based on instructors' expertise, but without much guidance for or contribution to theory.

Which of these approaches should cognitive scientists and educational researchers take to best address the problem of improving students' statistical reasoning abilities? (Hint: Think back to the story about the general!) The response advocated here is analogous to the general's solution: Do not use one or another solution approach by itself, but rather employ a *convergence* of multiple approaches. This chapter begins with the story about the general not as part of a

memory experiment (are you surprised?) but rather as a “source” problem to be applied by analogy to help solve the problem of improving statistics instruction (see Gick & Holyoak, 1983, for more details on their analogical problem-solving experiment).

While it is difficult to take a multidisciplinary approach to any research problem, the work of Ann Brown shows that it is possible and indeed can lead to striking benefits. In Brown’s 1992 “Design Experiments” paper, she emphasizes the importance of bringing multiple perspectives to bear for the success of her work. Brown’s training in experimental psychology spurred her initial theoretical ideas, based on memory and learning research, regarding how to improve instruction. These ideas led to instructional interventions (e.g., Palinscar & Brown, 1986) that she and her colleagues tested in the laboratory and then in the classroom, where they have become fairly widely used. More recently, she extended her applied side even further by playing the role of “learning community” builder in order to address issues of acceptance and impact so that her instructional innovations could produce larger educational changes. This integration of pure and applied work is a model for the success of the convergent approach to be discussed in the context of statistics education below.

The next section defines the term “statistical reasoning”. Then, a brief historical review of cognitive science research on statistical reasoning is

presented, highlighting how the three approaches mentioned above map onto the development of the field over the past 30 years:

1970's Cognitive science theories were developed to explain fallacies in people's statistical reasoning.

1980's New empirical work was conducted to study students' statistical reasoning

1990's Students' difficulties in statistical reasoning (and associated instructional innovations) were studied in the context of classroom practice.

These examples of single-perspective research are then followed by a description of some of my own research in which the aim is to bring these different solution approaches together to improve students' ability to transfer their statistical reasoning skills.

What is Statistical Reasoning?

"Statistical reasoning" is the use of statistical tools and concepts (e.g., hypothesis testing, variation, correlation) to summarize, make predictions about, and draw conclusions from data. Two examples illustrate this definition in a more probabilistic and a more statistical problem, respectively.

Example #1:

> A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital

> about 15 babies are born each day. As you know, about 50 percent of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower.
>
> For a period of one year, each hospital recorded the days on which more than 60 percent of the babies born were boys. Which hospital do you think recorded more such days? The larger hospital; the smaller hospital; or about the same (i.e., the two hospitals are within 5 percent of each other). (Kahneman & Tversky, 1972)

Solution to Example #1:

> The smaller hospital likely recorded more days with more than 60% boys because when sample sizes are smaller the observed distributions are more likely to diverge from that of the population at large.

Solving this problem requires reasoning about the impact of sample sizes on variability and involves making a prediction based on this relationship. Thus, this example fits the definition of statistical reasoning because (a) it employs the use of statistical concepts (e.g., sample size and variability) and (b) it involves making

a prediction based on the given data (two separate samples with sample sizes equal to 15 and 45).

Example #2:

> A weather modification experiment was conducted in Florida to investigate whether "seeding" clouds with silver nitrate would increase the amount of rainfall. Clouds were randomly assigned to be seeded or not to be seeded, and data were collected on the total rain volume falling from each cloud. A variable named *group* contains data on whether each cloud was seeded or not, and a variable named *rain* contains data on each cloud's rain volume. Does cloud seeding increase rainfall? To answer this question, perform any appropriate statistical analyses of the given data set and interpret the results accordingly.

(Partial) Solution to Example #2:

[Insert Figure 1 about here]

Solving example #2 requires performing exploratory and confirmatory data analyses. It thus fits the above definition of statistical reasoning because (a) it employs the use of several statistical tools and concepts (e.g., boxplots, confidence intervals, hypothesis testing) and (b) it involves drawing conclusions based on these analyses. It is important to note that the complete solution to this problem would involve more than the displays and statistics presented in Figure 1.

For example, the skewness of the data suggests re-analyzing the rain volume data after performing a transformation, an analysis not presented in Figure 1. More importantly, a complete solution would include a more thorough interpretation of the various results in terms of the question at hand, "Does cloud seeding increase rainfall?" (Note that this is a test problem used in the statistics course that provides a classroom context for much of my work described below.)

Historical Review

The question motivating this volume (and the Carnegie Symposium that it documents) asks how far we have come in applying cognitive research to instruction in the past 25 years. In this spirit, it is worth considering the development of past research relevant to statistical reasoning as a benchmark for current work in this area. Moreover, by looking at the past work on statistical reasoning—theoretical in the 1970's, empirical in the 1980's, and classroom-based in the 1990's—one can see how these different research approaches perform at their best in isolation, thereby gaining insights into how they might be profitably integrated.

A Template for Comparison

Four questions serve to structure this historical review of cognitive science research on statistical reasoning, The corresponding answers highlight how the social context of the research in each era influenced the nature of the research. A discussion of each of these questions follows:

Who are the researchers? The answer to this question may be cognitive scientists, education researchers, teachers, domain experts, or some combination of the above. It is an important question to consider because the background and experience of the researchers involved can greatly influence the direction of the research, both in terms of the questions being asked and the methods used to address them. In terms of identifying research that takes a "convergence" approach to studying statistical reasoning, the main distinction to be made regarding this question is whether or not the research integrates expertise from multiple disciplines (regardless of whether that multidisciplinary expertise comes from one or more than one person).

What is the goal of the research? The various answers to this question are of course quite involved, but they may be categorized into different types of goals, such as developing or extending a body of theory, adding to an empirical database, testing an educational intervention, or assessing educational outcomes (at a classroom or institutional level). While the goal of a particular project may not fall completely into one of these categories, it is likely that one emerges as most representative of the emphases in that project.

What is the context of the research? This question can be answered at many levels, as there are many ways of interpreting the term "context". For simplicity, a basic interpretation answers the question in terms of the *physical context* in which the research is pursued. Even at this basic level, there are vastly different

possibilities, including the cognitive psychologist's laboratory, the school classroom, or everyday life situations. Each of these different physical contexts tends to have a unique social and cultural context that can, in turn, influence the nature of the research.

How is the research being applied? The most salient application area is education and, more specifically, the improvement of instructional practice. This application may be a direct part of the work or a potential future outlet for the work. Alternatively, the research may not be directed toward any explicit application, or its application may be mostly self-directed, that is to extend or test a current theory. This question, then, does not address the nature of the work itself as much as the implications of the research, for example what impact will it have on science or society? and what audience will be concerned with its results?

With these framing questions outlined, the next three subsections briefly sketch research on statistical reasoning produced in three different periods: the 1970's, the 1980's and the 1990's. The purpose of this review is not to provide complete coverage but to convey the general tenor of the work produced in each of the different periods. For each period, the above questions will be answered based both on "typical" work and on specific papers from that era.

Theoretical Focus of the 1970's

Cognitive science in the 1970's savored the somewhat-new concept that mental representations can offer important insights into human behavior. Applying this

concept to the domain of statistical reasoning—specifically, to the question of how people make judgments under uncertainty—led to a body of work that is best represented by Kahneman, Slovic, and Tversky's (1982) book, entitled *Judgment under Uncertainty: Heuristics and Biases*. In this book, the editors, along with many other cognitive psychologists and social scientists working at that time, document their research from the preceding decade. They present a variety of behaviors that people exhibit when reasoning about probabilistic and statistical situations, and they posit specific mental processes as leading to these behaviors. In most cases, the behaviors demonstrate people's *errors* in reasoning. For example, the hospital babies problem presented earlier (Example #1) is taken from a study performed by Kahneman and Tversky (1972) in which more than half of the participants answered that the two hospitals would have the same number of days with more than 60% baby boys (despite the fact that one hospital delivers three times as many babies per day as the other). This answer fails to take into account the relationship between sample size and variability (i.e., the probability of extreme measurements), so it highlights the fact that people do not always employ correct statistical reasoning.

Other statistical reasoning fallacies (e.g., the gambler's fallacy, the conjunction fallacy, overconfidence, and insensitivity to base rates) were documented by these researchers. However, the main goal of the work was to develop and test a theory that explains *why* people make these reasoning errors.

Relying on the importance of mental representations and processes in understanding behavior, this theory proposed that people reason about uncertain situations by applying particular heuristics and biases that generally work well (i.e., produce "good" reasoning) but fail under particular circumstances. Thus, it should be emphasized that this era evince a positivist perspective, comparing human performance against logical ideals without much attention to the general value of using heuristics and biases (e.g., they are easy to apply, not overly computation- or memory-intensive, and make accurate predictions in many situations).

An example of one of these heuristics, proposed by Kahneman and Tversky, is the *representativeness heuristic*, which states that people make relative probability judgments about events A and B according to how representative A is of B. In the hospital problem, this heuristic applies when people consider the bigger and smaller hospital to be equally representative of the population of baby-delivering hospitals and therefore judge them both to have the same probability of delivering more than 60 percent baby boys on a given day. The representativeness heuristic explains several other errors in statistical reasoning as well. For example, Kahneman and Tversky (1973) showed that, in particular circumstances, individuals are insensitive to the prior probabilities of two events and that this insensitivity can be explained by application of the representativeness heuristic. When given brief personality descriptions of several

individuals taken from a pool of 70% engineers and 30% lawyers, participants rated the following individual as equally likely to be from either profession:

> Dick is a 30 year old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues. (Kahneman & Tversky, 1973)

Because this description was designed to convey no information particular to Dick's profession, the prior probabilities of the two professions would suggest that Dick was more likely an engineer than a lawyer. However, participants on average judged this probability to be 0.5. Kahneman and Tversky argued that this response stems from the representativeness heuristic because Dick's description is equally representative of a lawyer and an engineer, so the probability judgment is reasoned to be that he is equally likely to be one or the other.

Table 1 encapsulates the 1970's research on statistical reasoning by answering the four framing questions of this historical review. In summary, the research of this era mainly followed a theory-based approach to try to understand how people solve (or fail to solve) statistical reasoning problems. Little attention was paid to the realities of everyday problem solving, let alone classroom learning, and yet several new ideas and constructs were developed (e.g., reasoning heuristics) that could later be applied in more practical settings.

[Insert Table 1 about here]

Empirical Focus of the 1980's

Toward the end of the 1970's, researchers interested in science and math education were starting to get excited about the ways in which cognitive science theories and results (similar to those described above) could be applied to improve instruction (cf. Klahr, 1976). However, this potential influence between cognition and instruction did not immediately impact the domain of statistical reasoning. It was not until the 1980's that several researchers interested in statistical reasoning started working on the question of how *students* , who have been trained in a probability and statistics course, fare on similar tests of statistical reasoning (e.g., Fong, Krantz, & Nisbett, 1986; Konold, Pollatsek, Well, Lohmeier 1993; Pollatsek, Konold, Well, & Lima, 1984; Pollatsek, Well, Konold, Hardiman, 1987). These researchers tested students on problems that resembled those from Kahneman and Tversky's work. Presumably, a reasonable prediction at the time was that these students, thanks to their coursework in probability and statistics, would demonstrate better statistical reasoning than that exhibited in the earlier studies where the participants had no special training. However, this prediction did not generally hold true. For example, Fong et al. (1986) found that students had significant difficulty answering such questions outside the classroom context. Indeed, this experiment warrants special attention not only for its results on students' statistical reasoning but for its serious attention to the issue of transfer of relevant knowledge and skills outside the statistics classroom.

In Experiment 4, Fong et al. (1986) selected students at random from a college-level introductory statistics course; half of the students were tested during the first week of the semester, and the other half were tested during the last week of the semester. The "test", however, did not consist of a typical set of questions given in connection with the students' statistics course. Instead, students were contacted by phone (completely outside the context of any course) and asked if they would have time to answer some questions for a campus survey on students' opinions about sports. Following two questions that actually did ask students about their opinions regarding sports issues, four statistical questions—couched in terms of sports cover stories—were asked. An example of such a question that was designed to tap students' understanding of the concept of "regression to the mean" is as follows:

> In general, the major league baseball player who wins Rookie of the Year does not perform as well in his second year. This is clear in major league baseball in the past 10 years. In the American League, eight Rookies of the Year have done worse in their second year; only two have done better. In the National League, the Rookie of the Year has done worse the second year 9 times out of 10. Why do you suppose the Rookie of the Year tends not to do as well his second year? (p. 279)

Other statistical questions included in the survey involved the statistical concepts of sample size (cf. Example #1 presented earlier) and confounding variables. Students' responses to the four statistical questions were rated for evidence of statistical thinking and for the quality of the statistical response. For two of the questions, there was no significant difference between the two test periods (before and after students had taken a relevant course) in their use of statistical thinking. For the other two questions, there was a significant difference in the use of statistical thinking between the two testing times. However, in each of these latter two cases, the effect reflected only an additional 20% of students giving statistically related responses (with little concern of ceiling effects limiting the possible improvement). For example, for the question given above, only 37% of students tested at the end of the term used statistical reasoning in their answers, compared to 16% of students tested at the beginning.

Regarding the *quality* of statistical reasoning, only one out of the four questions showed a significant increase between students' responses collected at the two testing times. While these results do suggest that a statistics course can, for particular questions, produce transfer effects in students' statistical reasoning, an interpretation of the practical significance of these results (and similar results in Experiment 3 of the same paper) seems more akin to a glass half-empty.

Placing this work in its historical context, it is important to note that the motivation behind these experiments was not to understand students' statistical

reasoning in their college courses but rather to test the hypothesis that people in general tend to use abstract rules in their application of statistical concepts, such as the law of large numbers. This theoretical basis for studying students' statistical reasoning is also true of other work in this period. For example, a series of studies by Konold, Pollatsek, Well, and their colleagues was aimed at better understanding *why* students might demonstrate the various fallacies documented in the 1970's (Konold, et al., 1993; Pollatsek, et al., 1984; 1987). One such study tested two alternative hypotheses for why students reason incorrectly about random sampling: is it because they hold a passive, descriptive view of sampling that merely does not include the notion of independence of trials or is it because they hold an active-balancing model in which earlier trials influence later trials thus countermanding the notion of independence (Pollatsek, et al., 1984)? By varying the original paradigm used by Tversky and Kahneman, these researchers were able to distinguish that students' judgments were not consistent with an active-balancing model, thereby suggesting a descriptive view of sampling. These results helped to distinguish different underlying statistical conceptions that students might have, thereby expanding the 1970's theoretical work on reasoning fallacies; they were not, however, directly applied to the improvement of classroom instruction.

In summary, the research conducted during the 1980's, like that of the 1970's, was mainly theoretical in nature. However, there was a new trend toward

studying statistical reasoning in more realistic situations. Moreover, the research participants were often college students who had taken a course in probability and/or statistics. In this way, the results were more relevant to an applied audience including university instructors and instructional designers. Again, Table 1 summarizes the answers to the four framing questions for this era.

Classroom focus of the 1990's

By the 1990's, the reform movement in math and science education had taken a foothold in statistics instruction. Many new textbooks focusing on the practice of statistics were being used in courses whose curricula now emphasized reasoning about data rather than memorizing formulas. Along with this change in content, there was a change in the techniques of instruction: Students were getting more hands-on practice (typically in computer laboratories where they used statistical software packages to analyze data); they were solving more real-world problems; and they were getting access to computer simulations of various statistical phenomena (e.g., the central limit theorem).

With all these changes in statistics instruction, many researchers were asking the question: What do our students know and what problems can they solve coming out of these new courses? The motivation seems not to have been to compare previous, traditional instruction with the new, nor to directly explore the application of previous research to college instruction, but rather to document students' strengths and weaknesses under the reformed courses and, in some

cases, to evaluate the potential of specific instructional innovations. For example, Garfield and delMas (1991) tested students' conceptions of probability at the start and end of an introductory course by asking questions such as the following: Which of several coin flip sequences is most likely? least likely? How should one interpret a medicine label warning that there is a 15% chance of developing a rash? Each question had a set of multiple-choice answers. Comparing students' results from before and after the course showed that students did show an overall increase in correct responses. Nevertheless, particular misconceptions were maintained after instruction, leaving the absolute performance levels at posttest well below perfect performance. In a similar vein, Melvin and Huff (1992) listed and described several difficulties that their students demonstrated regarding various statistical concepts required for analyzing data and interpreting statistical results. In both of these studies, the results highlighted that students have difficulty applying particular statistical concepts even in the same context where these concepts were learned (i.e., not a case of real transfer).

Other work during this period shares the focus on assessing students' strengths and weaknesses but does so in the context of evaluating a particular new teaching intervention. For example, Cohen and his colleagues conducted several studies in which the students were tested before and after taking a statistics course that either did or did not employ a new instructional software package (Cohen and Checile 1997; Cohen, Checile, Burns, and Tsai 1996; Cohen, Tsai, and Chechile

1995). The questions were designed to test students' ability to apply the statistical concepts taught by the software package. Although the "experimental" students exhibited greater learning gains (posttest – pretest) than did "control" students, Cohen and Checile (1997) remark that "even those students with adequate basic mathematical skills [who had used the hands-on instructional software] still scored only an average of 57% [correct] on the [post-] test of conceptual understanding" (p. 110). While this result demonstrates an improvement relative to that group's average pre-test score of 42% correct, it shows that students' ability to reason statistically could still be greatly improved.

To summarize the research conducted during the 1990's, the main focus was on studying students' statistical reasoning (and difficulties thereof) in the classroom. (See Table 1 for the 1990's answers to the framing questions.) While there was little attempt to draw on previous theoretical work that might have helped to explain why these difficulties arose, there was a solid contribution in practical knowledge for teachers regarding where students' difficulties lie. Perhaps even more importantly, research in this era provided public documentation that there is still ample room for improving statistics instruction. It is this last point that may have provided an impetus for researchers to look to multiple perspectives in trying to make headway against the challenging problem of improving students' statistical reasoning.

Islands of Integrative Research in the mid 1990's and Beyond

As described above, most of the work on statistical reasoning during the 1970's, 1980's, and 1990's employed a single research approach (i.e., theoretical, empirical, or classroom-based). This past research also emphasized the products of learning and reasoning rather than the processes. There have been some recent studies, however, in which the focus is more process oriented and multiple approaches have been integrated. These bridging examples generally use the results of one approach to motivate or justify new research using a second approach. For example, Garfield (1995) provides a review of past results on probabilistic reasoning fallacies (highlighting some of the same 1970's research described above). She describes results, which were generated to test a more general psychological theory, but she does so in a way that draws on her own applied perspective and that will reach a classroom-oriented audience (e.g., high school and college statistics instructors).

Another paper by Garfield (1994) describes a fairly traditional assessment of students' statistical reasoning ability but adds to this applied topic a more theoretical perspective on learning. In particular, the assessment asks students not only for their answer to each question, but also which of several different modes of reasoning led them to their chosen answer. This approach enables the teacher/assessor to identify not only when and how often a student answers correctly/incorrectly but what mental processes and concepts led them to their answers. These additional data led Garfield to look for patterns of responding

across related questions to make better inferences about what mental models of statistics each student might have. More generally, it suggests to instructors and assessors to consider, in the abstract, what knowledge is required for statistical reasoning and how students' knowledge differs.

This approach is closely related to Hunt and Minstrell's DIAGNOSER research (e.g., Hunt & Minstrell, 1994) in which students are asked to complete assessments that require them to select among options that reflect both their solutions and their reasoning (i.e., both the products and processes of their problem solving). In this way, the assessment makes it possible to diagnose which *facets* of knowledge and understanding students have. Thus, like the theoretically focused work of the 1970's, this research aims to make inferences about students' internal mental states. In Minstrell and Hunt's work, however, the DIAGNOSER goes beyond inference to intervention in students' learning by providing feedback tailored to students' particular choices and asking additional questions tailored to the inferred facets of knowledge. This research integrates a theoretical and applied approach because it includes both the development of new theoretical constructs (i.e., *facets* of knowledge that represent certain profiles of understanding) and the development of an empirical database describing real students' various levels of understanding. Although Hunt and Minstrell's earlier work focused on students' understanding of physics, their approach has more recently been applied to the domain of probability (Schaffner, 1997).

A Convergent Assault on Statistical Reasoning

The remainder of the chapter will describe some of my recent work on examining Carnegie Mellon students' learning of statistical reasoning. This work is part of a project aimed at both *understanding* students' learning processes and *improving* their statistical reasoning abilities by creating new instructional environments. Exploratory and inferential data analysis (like that exemplified in the cloud-seeing example presented earlier) are the focus of our work. Our approach involves using cognitive theory to help achieve instructional goals, instructional results to help inform theory, and technology to help both. We are integrating all three approaches mentioned earlier: theoretical, empirical, and classroom based. In contrast with the work of the 1970's, 1980's, and 1990's, we are able to integrate these approaches in large part because of the multidisciplinary team working on this project. In our work, the researchers consist of cognitive psychologists, statistics instructors, educational researchers, *and* instructional technologists.

Our convergent approach is also made possible by the varied contexts in which we are conducting the research. These diverse contexts reflect a mixture of the psychology laboratory and the statistics classroom. In particular, the classroom context for our work is an introductory statistics class taken by more than 250 first-year undergraduates each semester. These students come from a variety of majors in the humanities, social sciences, and architecture.

Because of this strong link between our research program and the course mentioned above, it is important to briefly describe the content and format of the course. The instructors' goals in designing this course were to help students learn to (1) apply the techniques of exploratory data analysis, (2) understand the concept of sampling variability, (3) critically evaluate the effectiveness of different study designs, and (4) use and interpret inferential statistical procedures. (As mentioned above, exploratory and inferential analyses—goals 1 and 4—are the focus of our work.) The course de-emphasizes the memorization of statistical formulas in favor of students practicing statistics in authentic situations. Specifically, the students analyze real data sets in order to address current scientific and policy-oriented questions (e.g., Does seeding clouds with silver nitrate increase rainfall? Which of two new drugs is most effective in reducing the recurrence of depression? Do female professors earn less than their male counterparts?).

The format of the course includes two hour-long lectures and one hour-long computer laboratory session per week. In each laboratory session, students work in pairs at a computer, using a commercially available statistics package (Minitab 1999) to complete assigned exercises. These exercises are presented in the form of a lab handout that describes a data set, provides detailed instructions to guide students through the analysis, and asks them to interpret the results of their

analysis. Similar exercises are assigned as homework, where students apply the same skills without the supervision of laboratory instructors.

As it stands, this course uses several innovative instructional design techniques (e.g., collaborative learning, hands-on practice). However, given the 1990's research on similar reform-based courses, it seemed quite possible that students could be exiting the course without the desired statistical reasoning skills and transfer abilities. To address these potential areas of poor learning and transfer, we could have jumped in with a variety of new, "better" ideas to test in the class. Unfortunately, however, the existing research on students' difficulties in learning statistical reasoning does not offer much explanation of what causes these difficulties nor does it provide much guidance in devising specific solutions for overcoming them.

Thus, instead of relying on our intuition to guide us where past research could not, we took a systematic approach toward describing and understanding students' learning processes before we began any instructional interventions. Our approach involves developing a model of how students learn statistical reasoning, testing that model empirically, and using that model to inform instructional innovations. The products of our approach include: (1) a cognitive model of statistical reasoning that is detailed enough to solve the same problems that students will be asked to solve in an instructional setting, (2) well tested instructional innovations,

and (3) a computerized learning environment, based on (1) and (2), for students to use in the statistical reasoning class.

The remainder of this section documents four studies we have conducted that exemplify (1) and (2). Then, the following section describes how the results of these studies have been instrumental in our ongoing development of (3), a new computerized learning environment for statistical reasoning.

Task Analysis

Analyzing the knowledge and skills required for reasoning statistically is an important first step both for understanding how this skill is learned and for designing a learning environment to help teach it. Cognitive theory provides a mechanism for representing knowledge and can make detailed predictions about how that knowledge is learned and used. Within the ACT-R cognitive theory (Anderson & Lebiere, 1998), knowledge is represented in one of two ways: as declarative facts arranged in a semantic network or as procedural skills embodied in a set of production rules, each of which specifies an action to be taken under particular circumstances. These pieces of knowledge can be inferred from traces that document the steps a solver takes at each point in a problem. Once the set of production rules and semantic network are specified, the corresponding problem-solving performance can be generated by simulation using the ACT-R computational engine. Putting these knowledge pieces into a cognitive model

makes it possible to compare the theory's predictions with observed behaviors and to evaluate or refine the model.

We began this process of knowledge decomposition by focusing on the first and last of the four course goals that were specified by the instructor (as listed above). These goals refer to students' ability to apply the techniques of exploratory data analysis and inferential statistics. By collecting talk-aloud protocols (Ericsson & Simon, 1993) from the statistics instructors and analyzing what they considered to be "ideal" student solutions to typical data-analysis problems, we generated a sequence of steps that reflect their problem-solving process (See Figure 2). Note that this sequence of steps can be applied to solve the cloud-seeding example presented earlier. To refine this analysis and obtain a further specification of each step, we generated an initial set of production rules and corresponding declarative facts that enabled the ACT-R simulation program to perform each of the steps (correctly and at the appropriate points in problem solving). Testing the model on sample problems assigned in the course, comparing the model's performance to that of actual students', and identifying parts of problems where the model lacked appropriate knowledge, we iteratively refined our cognitive model. The model thus represented (to a reasonable level of completeness) the facts and skills that students would need in order to solve each part of a typical problem.

[Insert Figure 2 about here]

Given our focus on improving students' ability to transfer their statistical reasoning ability, we were interested in how much of this knowledge would be general (i.e., applicable in multiple contexts for multiple problems) and how much of it would be case specific (i.e., only applicable under specific circumstances). The model was helpful in making this distinction because it highlighted the fact that there were particular production rules that were used regardless of the specifics of the problem. For example, Table 2 shows several of these production rules, translated to pseudo-code for easier reading. The first production rule represents the knowledge for executing a particular step, namely, choosing an appropriate graph. Notice, however, that even this particular graph-choice step is represented with sufficient generality that it applies to problems with different types of variables. The second and third production rules in the table represent a small part of the common goal structure that guides the application of particular problem-solving steps throughout the solution. This goal structure is depicted at a high level in Figure 2 and, like that figure, applies to almost any exploratory data analysis problem.

[Insert Table 2 about here]

For the purpose of understanding students' learning processes, this production representation raised the question: At what level of generality do students learn these rules. For example, students could learn lots of specific rules to apply in different situations instead of the more general rules in our model, or

they could learn overly general rules that do not adequately account for the specific conditions under which particular steps are appropriate. For the purpose of improving students' learning via our new learning environment, this question led to the further consideration of how we could design the learning environment to emphasize the common goal structure so that students' internal knowledge representations would be at an appropriate level of generality, like the model's (see the next major section for more details on this issue).

This analysis and model development were also quite helpful in establishing a common language among the statistics instructors and cognitive psychologists that facilitated the collaboration in our group. As we developed the model and discussed its possible refinements, the instructors had opportunities to indicate precisely which aspects of the problem-solving process they found important, and the cognitive psychologists had concrete, domain-specific examples to use in discussing how knowledge comes in different forms (e.g., procedural vs. declarative) and the implications of different levels of generality.

Large-Scale Assessments

Having generated a model describing what students need to know to reason statistically on data analysis problems, we next wanted an idea of students' knowledge, both before and after taking the course. This information would be especially helpful for our applied goal of improving instruction because it would indicate students' strengths and weaknesses, that is, where we needed to work

most to improve their statistical reasoning abilities. In addition, because of our focus on transfer, we wanted some assessment of how students would perform outside the original learning context. To obtain this information, we created an assessment instrument designed to tap students' statistical reasoning abilities. In particular, we used the results of our task analysis to generate separate questions for each separate concept (a part of the model's semantic network) or skill (a subset of the model's production-rule set). In this way, we tried to achieve a close correspondence between items on the instrument and particular pieces of knowledge identified in our task analysis as critical for statistical reasoning.

For ease of administration, scoring, and analysis with large groups of students, we designed the questions in multiple-choice format. While this conveys more information about the products of problem solving (which is different from most of our other work which emphasizes process), the fact that the questions focus on small sub-parts of each problem gives some intermediate information about students' problem-solving process. As examples of the content and format of the assessment, Figure 3 shows two questions that were designed to test students' knowledge of how to choose an appropriate graph.

[Insert Figure 3 about here]

We administered the complete assessment according to a pretest/posttest design with two groups of students participating—those who did and those who did not take the course between the pretest and posttest phases. The latter group

served as a control so that we could determine what statistical reasoning abilities students in the course were gaining above and beyond those attributable to maturation or exposure to the general college environment. Both groups of students took the pretest in an extracurricular testing session given at the beginning of the school year. Then, at the beginning of the second semester of that same year, students took the posttest in the context of a separate statistics course.

This posttest context represents our attempt at establishing a near-transfer situation: "near" because the testing environment was another statistics course (as opposed to, say, an economics course) but still "transfer" because that course was different from students' original learning environment. Given poor results in the studies reviewed above, it seemed prudent to test for near transfer before expanding to more distant transfer tests. Note that the context of this posttest also had important ecological validity that was of interest to the statistics instructors; we were testing students' ability to retain and apply what they learned in one course to a related course downstream in their program of study.

Here, we summarize the results of this assessment and describe how they were helpful to the different members of our group. (Lovett, Greenhouse, & Johnson, 1999, provides a more thorough report of our analyses of the data.) Panel a of Figure 4 presents total percentage correct for each group of students at pretest and posttest. These results were encouraging in that they showed that

students who took the statistics course improved their total scores more than did the students who had not taken the statistics course. Such aggregate results, however, do not provide any diagnosis of students' potential areas of strength and weakness. We therefore conducted item-based analyses of students' responses and found three distinct categories of items: (1) items for which the statistics-class students increased the accuracy of their responses but for which the other group of students did not, (2) items for which both groups of students showed no increase in accuracy but could have, and (3) items for which both groups of students showed no increase in accuracy due to a ceiling effect at the pretest. Panels b, c, and d of Figure 4 show the average proportion correct for group of students for each of these categories of items, respectively. Only the pattern of results in panel b demonstrates learning of statistical reasoning skills that can be attributed to the course.

[Insert Figure 4 about here]

By identifying which items on our test fall into these three categories, we were able to glean important information both about students' areas of strength and weakness (i.e., what components of statistical reasoning skill are inherently difficult?) and/or how the course might be improved (i.e., what part of the course is the instruction insufficient?). For example, the first category of items included skills such as interpreting descriptive and inferential statistics and defining statistical terms (e.g., correlation). These skills appear to be well learned from the

course, and there is not much need for improvement. The second category of items included skills such as choosing appropriate statistical displays (including the example questions from Figure 3) and drawing conclusions from statistical analyses. The corresponding pattern of performance in Figure 4c suggests that students have particular difficulty with these skills—both before and after the course—and that our efforts at improving the course should be directed at these areas. Finally, the third category of items included interpreting boxplots and scatterplots. That students had an especially easy time with these subskills was somewhat surprising to the instructors. This result implies that the course need not emphasize these subskills as much as it does, something that we could also take into account in our plan for a new learning environment.

Assessing students' abilities skill by skill with this instrument gave the instructors more precise information about the areas of strength and weakness than they could previously obtain. Final exam questions given by instructors often require synthesis of a variety of skills (an important ability to test) but do not offer the same diagnosis capability, nor do they offer a controlled comparison to students' abilities before the course or without having taken the course. In summary, these assessment results provided important information to the instructors about their students' strengths and weaknesses, and they provided initial pointers for the project team as a whole in terms of where we should

concentrate our further study of why students have difficulty reasoning statistically.

Detailed Study of Individual Students

As a result of the large-scale assessments, we knew that students exited the course without having fully learned particular subskills important to statistical reasoning (e.g., choosing appropriate analyses, drawing conclusions, etc.), but we did not know the source of these problems. Moreover, we wanted to know how students were able to combine these and their better learned skills in the context of solving authentic problems. To address these questions, we conducted a very different type of study, focusing on individual students' ability to apply statistical reasoning in solving open-ended problems. We asked individual students who had taken the statistics course under study to come into the psychology laboratory and provide talk-aloud protocols while they solved a few data analysis problems analogous to the ones they had encountered in class. They were allowed to use whatever statistics package they preferred (usually the one they had used in the statistics class). The main difference between these experimental sessions and students' computer laboratory sessions in the course is that our problems were stated with only three basic pieces of information: the background needed to understand the problem, the research question being asked, and a description of the data to be analyzed. In contrast, in most of the computer laboratories in the statistics course, students would have received this information plus an entire lab

handout guiding them explicitly as to how they should proceed in their problem solving. Instead of constraining students to a single solution path, we wanted to investigate how they would approach these problems in a much more open format, i.e., what ideas and strategies for data analysis that they would generate on their own. Moreover, this served as a different kind of transfer test to see how well students could solve problems without the typical aid of a lab handout to guide them.

We collected students' talk-aloud protocols and synchronized these with the computer traces of their interactions with the statistics package. Together, these two streams of data offered a rich description of the mental and physical steps students were taking as they solved the problems. Here is a sample problem from our study:

> In men's golf, professional players compete in either the regular tour (if they're under 51 years old) or in the senior tour (if they are 51 or older). Your friend wants to know if there is a difference in the amount of prize money won by the players in the 2 tours. This friend has recorded the prize money of the top 30 players in each tour. The variable *money* contains the money won by each of the players last year. The variable *tour* indicates which tour

the player competed in, 1=regular, 2=senior. The variable *rank* indicates player rank, 1=top in the tour.

We analyzed students' verbal and computer protocols in several ways to get a full description of their problem-solving behavior. First, we coded the combined protocols according to the main steps of problem solving (see Figure 2): If a given protocol segment or computer interaction offered evidence that the student had considered one of these steps, we would code the step as attempted. Figure 5 shows a protocol excerpt and our coding of each segment. The percentage of students who showed evidence of engaging in each step is presented in Figure 2 next to the box corresponding to that step. It is clear that students were often not engaging (at least explicitly) in the first three steps and the last step of our problem-solving sequence. While it is possible that experts could skip the first three steps and initiate their problem solving at step 4 (selecting the appropriate analysis), this sample of students did not demonstrate such expertise: Although 100% of protocol subjects showed evidence of attempting to select the appropriate analysis, their accuracy in doing so was only 50%. Thus, it seems likely that inaccuracy in step 4 was in part caused by skipping steps 1-3. Separate analyses supported this idea by demonstrating that the probability of a correct step 4 was much higher in cases when the preceding steps were not skipped compared to when they were skipped. Also, note that although 100% of subjects gave *some* interpretation of their results, only 80% provided accurate interpretations.

[Insert Figure 5 about here]

Given students' difficulty with step 4, choosing the appropriate analysis, our next step was to review the protocols to explore the nature of students' inappropriate choices. More specifically, what (if anything) were students doing instead of applying the preceding three steps that could lead them to an appropriate analysis? An "interpretation approach" (Chi, 1997) was used where the protocols were examined to facilitate interpretation of the computer data. In many of the verbal protocols, we found evidence that students were relying on the statistics package as a crutch to get a reasonable analysis on screen. Two such examples are presented below. In the first, the student does not systematically derive the appropriate analysis given the problem information but rather uses the statistics package's menu list as an idea generator:

> Oh, okay. Um, I'm not really sure if- do I need to uh we can just, like, graph it, right? Uh line plot, I guess. … oh, uh histograms, barcharts maybe a boxplot? Uh, no... Uh, uh histogram, um data table, um...

In this case, there is no clear constraint on the student's selection process, nor is it guided by a conceptual understanding of the task.

In the second protocol example, a different student uses two separate heuristics for selecting the appropriate analysis, neither of which is related to the

specifics of the problem or an understanding of the task. The first heuristic involves relying on what is typically a correct choice in this task (i.e., following the base rates of success on past problems). The second heuristic involves using the statistics package's warning message as feedback that the chosen analysis is not appropriate:

> Oh well, maybe, hmm... if I highlight all of them [the variables in the dataset], and then, maybe make a boxplot cause, in statistics class that always worked when you got stuck, just make a boxplot, and see what happened. So uh, I'll boxplot them, um, y by x. [quack] Uh oh, it says the variable rank has 30 categories, shall I continue? Usually that was bad, so I cancel that, because it shouldn't come out like that.

These protocol examples offer a preliminary hypothesis for why students were (1) skipping the first three planning steps of the problem-solving process and (2) relatively inaccurate in selecting an appropriate analysis. Namely, by using the statistics package interface cues, they were able to apply a basic guess-and-test strategy in order to generate analysis.

Our third analysis of the students' problem-solving traces looked for quantitative evidence supporting this hypothesis. In particular, we analyzed

various features of the analyses students performed on each problem. We found that, on average, students performed approximately 11 separate analyses per problem, even though only three analyses at most could be deemed truly appropriate. Also, of the analyses students generated in their problem solving, approximately three per problem were exact repeats of a previously generated (usually inappropriate) analysis. These two results suggest that students were not using an efficient search strategy because they were generating so many extra (generally useless and often redundant) analyses, Moreover, the sequence of analyses generated by the students did not follow what the course had taught. On average, the most informative statistical analysis (i.e., the one that the course instructor would have performed first and the one that was most consistent with the teaching in the course) was the *sixth* analysis attempted by these students.

There are at least two possible explanations for this pattern of results. One is that students have not yet learned a systematic procedure for selecting appropriate displays that works for all sorts of data analysis problems. Thus, they do not see the common structure across problems and do not know how to proceed in a systematic fashion. Another possible explanation is that students have arrived at a sub-optimal strategy that enables them to "get by" with arbitrary selections but without understanding of the reasoning behind their steps. Both of these suggest that our computerized learning environment should emphasize the

common goals and procedures across problems and monitor students' choices to assess the effectiveness of their selection strategies.

Experimental Study of Learning

The above results helped us uncover an important area of difficulty in students' statistical reasoning—the ability to systematically plan an analysis based on the problem information and on an understanding of statistical displays. Tackling this difficulty area thus became one of our new goals. In particular, we wanted to design our computerized learning environment to facilitate students' planning processes so they could more easily learn to choose appropriate statistical displays and acquire the corresponding skills at an appropriate level of generality. We generated various ideas, based on past research, that would *scaffold* students in this planning process. Before implementing any of our ideas in the context of a full-scale learning environment, however, we compared two potential design variants in a controlled experiment. Our motivation for doing so was twofold. First, as a basic research goal, we wanted to gather more, fine-grained data on how students' *learn* these planning skills. Note that the studies presented above involved either students who had already taken the statistics course (i.e., were not in an initial learning phase) or a data-collection procedure that produced very coarse-grained information (i.e., provided data only on students' answers not their processes). In the description below, we describe data collected at a fine grain from students who had not taken a previous statistics course. Our second

motivation was an applied research goal. We wanted to gather some preliminary data on whether our ideas for scaffolding students' planning skills would actually aid learning. In particular, we wanted to compare two versions of a computer interface which manipulated the degree of scaffolding students would receive as they learned how to choose appropriate statistical displays.

The procedure of this experiment involved assigning students to one of two conditions (i.e., end-only and immediate feedback) and then asking them to complete four experimental phases (i.e., pretest, instruction, problem solving, and posttest). The first phase involved a set of pretests to assess students' pre-experimental understanding of statistical displays and planning. These tests included problems where students had to consider the entire problem-solving situation, not just the step of choosing appropriate displays. The second phase involved an instructional phase in which students read various materials (on the computer) that described different statistical displays, how they are produced and under what conditions they are appropriate. These materials were made available to students throughout the course of the experiment, whenever students chose to access them. The third phase, the only phase to differ between the two groups, involved a series of 16 problems that students were asked to solve on the computer. Depending on their assigned condition, students would receive more or less specific feedback as they worked through each problem. Figure 6 shows the problem-solving interface. Students in the "end-only" scaffolding condition

were asked to make all five selections shown, in any order, and then submit their answer. Upon doing so, they were shown the corresponding statistical display (regardless of whether their selections were correct) and binary feedback (correct/incorrect) regarding their entire response. Note that incorrect feedback for the end-only group does not disambiguate which of the five choices is incorrect. In contrast, students in the intermediate-feedback condition received correct/incorrect feedback after making the first four choices (choosing the response/explanatory variables and their quantitative/qualitative type). If incorrect, they would be forced to try again until these choices were correct. The procedure thus implies that this group of students would only be selecting a type of display (i.e., making the fifth choice) after they had correctly classified the problem situation. Further, it implies that their end-of-problem feedback (same as in the other group) unambiguously referred to the correctness of the fifth (display type) choice. The fourth phase of the experiment involved the same tests used in the first phase.

[Insert Figure 6 about here]

The data gathered in this experiment consisted of students' answers to the pre- and post-experimental tests (phases one and four) and complete traces of their interactions with the computer during phases two and three. Figure 7A shows that, based on their pre/posttest scores, students improved a great deal in their ability to select appropriate data displays, $F(1,50)=69.6$, $MSE=2.83$, $p< .01$.

As a point of comparison, it is interesting to note that students in this experiment, who had taken no previous statistics class and who spent approximately 45 minutes working with these instructional materials and problems, showed posttest scores that were comparable to those of students who had taken the full-semester course and then participated in the same experiment. This comparison does not suggest that students can learn an entire statistics course in 45 minutes, but rather that the set of subskills involved in selecting appropriate analyses can be reasonably well learned in a short, focused lesson that forces students to practice making these selections on their own. It is often the case in the context of an actual statistics class that students do not actually have many opportunities to make such choices on their own; these choices are either made for them explicitly (in homework or lab assignments that indicate which analysis is appropriate and ask for students' interpretation of the results) or implicitly (in cases where there is only one new analysis being taught in a given week and that analysis is the correct one for the problems assigned that week).

[Insert Figure 7 about here]

Perhaps more interesting than the pre/posttest data are the data collected *while students were learning*. On average, the number of attempts made on each problem decreased with problem number for both conditions in the experiment, $F(3, 54) = 2.9$, $MSE = 13.6$, $p < .05$ (See Figure 7B). In other words, students were getting better at solving the problems over the course of the experiment.

Moreover, the number of attempts across blocks of the experimental phase was lower for the intermediate-feedback group, suggesting that this version of the interface made the learning process go more smoothly and quickly, $F(1, 18) = 3.9$, $MSE = 37.3$, $p < .06$ (See Figure 7B). The advantage of the intermediate-feedback group was also revealed when analyzing a particular set of "difficult" problems where it was predicted that students would tend to make errors: On these problems, the intermediate-feedback group chose the correct analysis first 82% of the time, whereas the end-only feedback group chose the correct analysis first 40% of the time.

These results support two general points about how students learn to choose appropriate statistical displays. First, students can acquire mastery of this skill by practicing it in isolation with adequate feedback. This supports the notion that decomposing the task of statistical reasoning into the required knowledge and skills for good performance can lead to targeted, effective instructional interventions. Of course, this knowledge-decomposition idea also acknowledges the value of giving students practice at the "synthesis" skills that are required for handling whole-problem solutions. Second, the fact that intermediate feedback helps students learn this skill more efficiently suggests that students can benefit from more than a standard statistical software package when learning. In particular, the "feedback" offered by a statistical software package is limited in that (1) it relies on the student's ability to interpret a dubious display as such, (2)

it does not indicate what aspect of the student's selection is incorrect, and (3) it does not provide any information to help the student to correct the error. In contrast, the intermediate feedback condition of this experiment provided enough information to avoid all three of these problems. In our learning environment, we are incorporating feedback features that avoid these problems as well as offering students information on why their selections were wrong.

A Convergent Assault: Putting It All Together

As the above studies show, understanding how students learn statistical reasoning can be studied effectively from many different perspectives. Moreover, greater gains can be achieved when these perspectives are brought together to influence each other. In each of the four studies described, multiple perspectives were integrated (e.g., theoretical and empirical, empirical and classroom-based, etc.) to study a particular aspect of students' statistical reasoning. These four studies also serve to provide important results that are informing the design of our computerized learning environment for statistical reasoning. In this way, all three perspectives converge to impact the way students learn statistical reasoning in the classroom.

The design considerations offered by the above studies' results are as follows. First, the task analysis highlighted the general skills that correspond to the goal structure present in many data-analysis problems. Ideally, then, our learning

environment should help students to learn these skills in a general way so that they can transfer what they learn to a variety of problem contexts. Second, the large-scale assessment indicated that, while students improve overall on statistical reasoning questions after taking a course, there are particular areas (e.g., selecting appropriate displays, and evaluating the strength of evidence) with ample room for improvement. Our learning environment should give special attention to these aspects of statistical reasoning, such as scaffolding students' intermediate steps. Third, the detailed, process-baed study showed that students' difficulties in planning stemmed from the application of non-optimal strategies for selecting appropriate analyses (e.g., guessing through menu items in the statistics package). Our learning environment thus should discourage students from using these strategies and instead should teach them to apply a systematic strategy based on an understanding of data types and experimental designs. Fourth, the laboratory study showed that practice on planning steps improves students' ability to select appropriate analyses and that intermediate feedback increases students' efficiency of learning.

Applying all these considerations jointly leads to the design of a learning environment that has the following features. First, to highlight the general schema for solving data-analysis problems, the learning environment should make the goal structure explicit. Other researchers have achieved this by labeling important

goals and subgoals in problem solutions (Catrambone 1995, 1996) and by emphasizing the commonalties across problems (Cummins, 1992). Figure 8 presents a snapshot of a prototype for the interface to our learning environment. Notice that the “outline-like” format presents major goals and subgoals with expand/compress buttons for focusing on particular parts of the problem. The labels for these goals and subgoals are the same for all data-analysis problems, regardless of the particulars of the dataset or questions. Second, our learning environment scaffolds students in their planning processes. The goal structure highlighted in the interface includes steps for considering the relevant variablesand their types that students must complete before selecting a particular analysis for the given problem. Here the aim is to reduce students’ “dive in” tendency and to encourage them to explicitly plan their analysis. This interface also makes the invisible skills of planning visible (cf. Koedinger & Anderson, 1993) by giving external actions for steps that ordinarily would only take place “in the students’ head”. These external actions then enable the third design consideration, offering feedback to students at critical points in problem solving. When studens can communicate their intermediate planning steps to the problem-solving interface in Figure 8, the problem-solving engine behind this interface can offer feedback on these steps individually. For example, if a student identifies the response variable incorrectly, that mistake can be indicated and explained *before* the student goes on to conduct and interpret analyses that have no relevance to the

question at hand. Note that the problem-solving engine here is based on the cognitive model of statistical reasoning developed in the task analysis study; it makes our learning environment an example of an intelligent tutoring system because it can track students' problem solving and offer hints and feedback accordingly.

[Insert Figure 8 about here]

As we develop and refine this learning environment, we will not abandon the convergent approach that motivated its development; our evaluation protocol for our own system will be based on empirical studies conducted both in the laboratory and the classroom, and we will use these results in combination with theoretical considerations and the guidance of our domain experts to improve the effectiveness of the system. Part of this further development will include going beyond exploratory and inferential data analysis to emphasize the other two goals of the course, namely, sampling variability and experimental design.

Making the General's Strategy Work

Although it was not discussed in the opening story about the General, an important pre-requisite for the success of his convergent strategy is effective communication. Coordinated timing was the key to the strategy; if communication among the various troops were not strong, the whole plan could have been dashed. Similarly, in the collaboration discussed in this chapter,

effective communication among the different team members has been critical. Such communication, however, takes time to establish in a multidisciplinary situation. Early on in the project, team members with different areas of expertise spoke somewhat different languages. It took collaboration on specific problems of mutual interest, where everyone was willing to consider alternative perspectives, to establish a common language. As this common language has been refined through our work on the project, the synergy of our multiple perspectives has increased.

References

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

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences, 2,* 141-178.

Catrambone, R. (1995). Aiding subgoal learning: Effects on transfer. Journal of Educational Psychology, 87(1), 5-17.

Catrambone, R. (1996). Generalizing solution procedures learned from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences, 5,* 3:271-315.

Cohen, S., & Chechile, R. A. (1997), Overview of ConStats and the ConStats Assessment, In J. B. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics*, Voorburg, The Netherlands: International Statistical Institute.

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, 35-54.

Cohen, S., Tsai, F., & Chechile, R. (1995), A model for assessing student interaction with educational software, *Behavior Research Methods, Instruments, and Computers,* 27, 251-256.

Cummins, D. D. (1992). Role of analogical reasoning in induction of problem categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1103-1138.

Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis: Verbal reports as data. (Revised Edition). Cambridge, MA: MIT Press.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18,* 253-292.

Garfield, J. (1994).Beyond testing and grading: Using assessment to improve student learning . *Journal of Statistics Education,* [On-line serial], *2*(1), Available E-mail: archive@jse.stat.ncsu.edu Message: Send jse/v2n1/garfield.

Garfield, J. (1995), How students learn statistics, *International Statistical Review,* 63, 25-34.

Garfield, J., & delMas, R. (1991), Students' conceptions of probability, In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (Volume 1), (pp. 340-349), Voorburg, The Netherlands: International Statistical Institute.

Gick, M., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1-38.

Hunt, E., & Minstrell, J. (1994). A cognitive approach to the teaching of physics. In K. McGilly (Ed.), *Classroom Lessons: Integrating cognitive theory and classroom practice.* (pp. 51-74). Cambridge, MA: MIT Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237-251.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* New York, NY: Cambridge University Press.

Klahr, D. (1976). *Cognition and Instruction.* Hillsdale, NJ: Erlbaum.

Koedinger, K. R., & Anderson, J. R. (1993). Reifying implicit planning in geometry: Guidelines for model-based intelligent tutoring system design. In S. P. Lajoie, & S. J. Derry (Eds.) *Computers as Cognitive Tools* (pp. 15-45). Hillsdale, NJ: Erlbaum.

Konold, C., Pollatsek, A., Well, A., Lohmeier, J. (1993). Inconsistencies in students' reasoning about probability. *Journal for Research in Mathematics Education, 24,* 392-414.

Lovett, M. C., Greenhouse, J. B., & Johnson, M. (1999). *Assessing an introductory statistics course.* In preparation.

Melvin, K. B., & Huff, K. R. (1992). Standard errors of statistics students. *Teaching of Psychology, 19*(3), 177-178.

Minitab (1999), Statistical Software (Student Version 9). Reading, MA: Addison-Wesley Publishing.

Palinscar, A. S., & Brown, A. L. (1986). Interactive teaching to promote independent learning from text. *Reading Teacher, 39,* 771-777.

Pollatsek, A., Konold, C. E., Well, A. D., & Lima, S. D. (1984). Beliefs underlying random sampling. *Memory & Cognition, 12,* 395-401.

Pollatsek, A., Well, A. D., Konold, C., Hardiman, P. (1987). Understanding conditional probabilities. *Organizational Behavior and Human Decision Processes, 40,* 255-269.

Schaffner, A. A. (1997). *Tools for the Advancement of Undergraduate Statistics Education.* University of Washington [Dissertation Abstracts, Vol. 58-08A, p. 3056].

Acknowledgements

The author would like to thank Joel Greenhouse, Brian Junker, Rob Kass, and Ken Koedinger who have been and continue to be key collaborators in this project. This work is partially funded by a grant from the National Science Foundation.

Table 1

Contrasting Historical Periods of Research on Adult Statistical Reasoning

Question	1970's Answer	1980's Answer	1990's Answer
Who are the researchers?	Cognitive psychologists and social scientists	Cognitive psychologist	Psychologists, educators, and instructors
What is the goal of the research?	Developing and testing a theory positing that people use certain heuristics and biases	Primarily testing theory but also documenting abilities of statistics students	Documenting students' difficulties in statistical reasoning
What is the context of the research?	The psychologist's laboratory, with sanitized versions of real-world problems	Studying students outside the context of statistics class, with pseudo-real-world problems	Studying students in the classroom
How is the research	To develop and test	To test theory; could	To provide

being applied?	theory	also be applied to instructional design.	information to instructors of similar courses

Table 2
General Production Rules in Cognitive Model of Statistical Reasoning

If the goal is to address question q with dataset d
& the relevant variables are x and y
& the type of x is x-type & x is the explanatory variable
& the type of y is y-type & y is the response variable
& a graphical tool for y-type versus x-type data is graph g
THEN produce a graph g of y versus x

IF the goal is to address question q with dataset d
& relevant variables have not yet been selected
& variables x and y are in dataset d and relevant to question q
THEN select x and y as the variables to be analyzed

IF the goal is to address question q with dataset d
& the relevant variables are x and y
& the type of variable x (quantitative/categorical) has not yet been identified
THEN set a subgoal to inspect variable x as to type

Figure Captions

Figure 1. Components of a solution to the cloud-seeding problem.

Figure 2. Task analysis of major steps in solving exploratory data analysis problems. Note that this is a cyclical process in which initial analyses may suggest further questions for analysis. Percentages to the right of each step represent the percentage of students in a protocol study who showed explicit evidence of engaging in that step.

Figure 3. Two sample assessment questions on choosing appropriate statistical display.

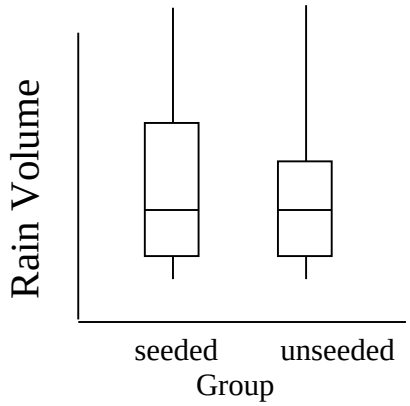
Figure 4. Results of large-scale assessment. Panel A shows percentage correct over all items. Panel B shows percentage correct for items on which students who took the course improved from pre- to posttest. Panel C shows percentage correct for items on which both groups of students showed no improvement, even though they could have. Panel D shows percentage correct for items on which both groups of students showed no improvement, presumably due to a ceiling effect at pretest.

Figure 5. Sample protocol with each step coded according to the major steps of statistical reasoning (see Figure 2). Notice that there is no evidence for steps 3 and 7. Also notice that the interpretation is somewhat inaccurate in that boxplots display the median not the mean as a measure of central tendency.

Figure 6. Interface displayed to subjects in learning experiment.

Figure 7. Panel A shows the overall improvement in pre/posttest scores for the two groups of student—those with and without previous statistics courses. Panel B shows the improvement throughout the course of the experiment for the two different conditions (end-only vs. intermediate feedback), including only those students who had no prior statistics classes.

Figure 8. A snapshot of a prototype interface to our learning environment.



	N	MEAN	MEDIAN	STDEV
Seeded	45	61.33	40	45.31
Unseeded	45	44.67	40	35.07

The exploratory analysis (i.e., boxplots and descriptive statistics) suggest that there is great overlap in the distributions of rainfall among the "seeded" and "unseeded" clouds. [Ideally, student would also notice skewness of rainfall data and hence would perform inferential statistics on transformed data.

95% Confidence interval for $\mu_{seeded} - \mu_{not}$: (0.02, 0.583)

T-test for H_0: $\mu_{seeded} = \mu_{not}$ H_A: $\mu_{seeded} \neq \mu_{not}$

t= 2.11 p=0.038 df= 85

These inferential statistics suggest that, with an alpha level of 0.05, these data show a significant evidence to reject the null hypothesis of no difference between seeded and unseeded clouds' rainfall volumes. This is consistent with the 95% confidence interval which falls completely above 0.0, suggesting that the seeded clouds' rainfall volume is greater than the unseeded clouds'.

Figure 1. Components of a solution to the cloud-seeding problem.

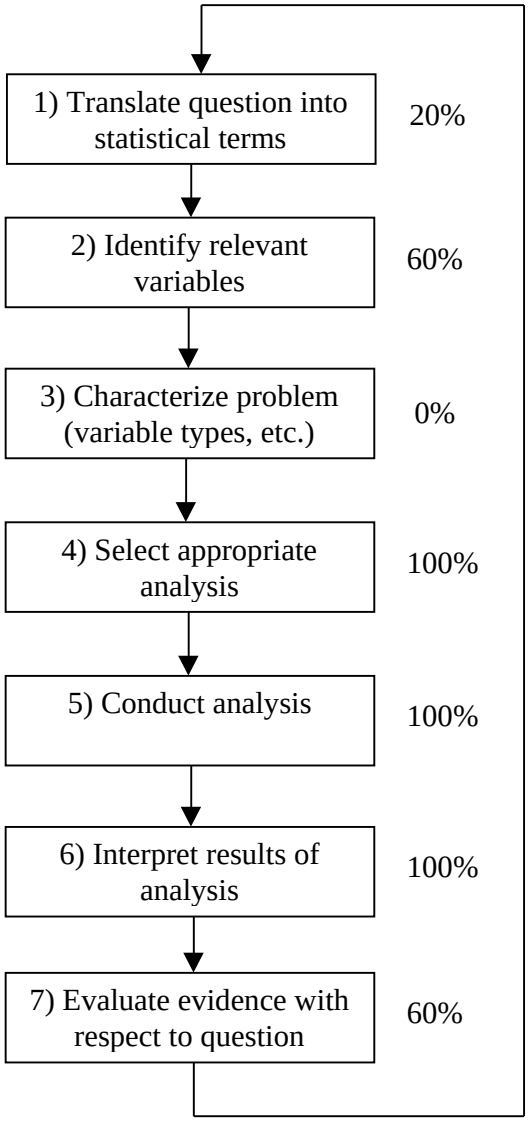
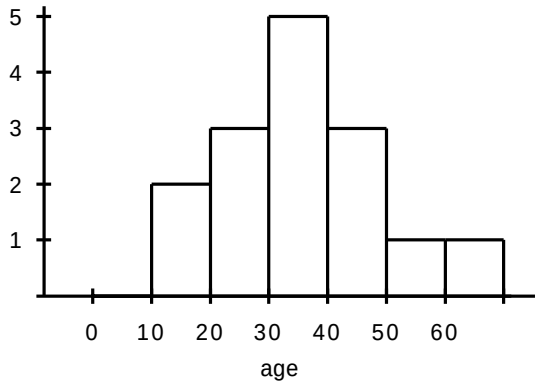


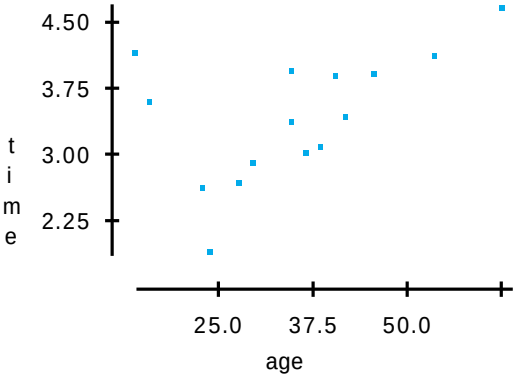
Figure 2. Task analysis of major steps in solving exploratory data analysis problems. Note that this is a cyclical process in which initial analyses may suggest further questions for analysis. Percentages to the right of each step represent the percentage of students in a protocol study who showed explicit evidence of engaging in that step.

Your local running club has its own track and keeps accurate records of each member's age and individual best lap time around the track, so members can make comparisons with their peers. Below are graphs of these data.

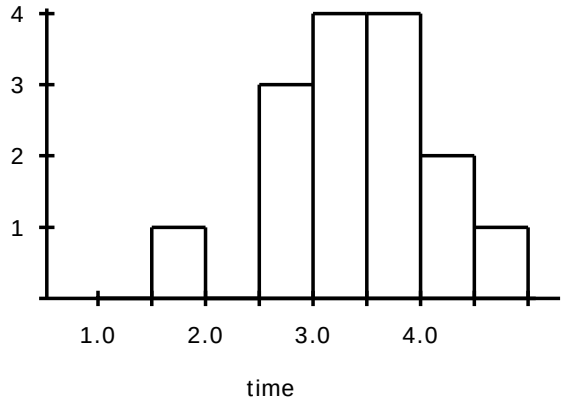
Graph A



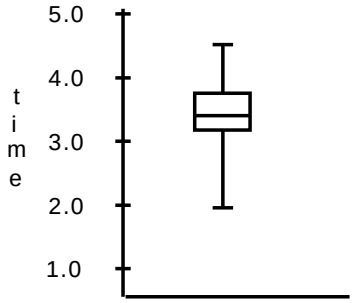
Graph B



Graph C



Graph D



12. Suppose you wanted to show a group of new members how much joining this team improves people's running times. Which of the four graphs above would you use? [Answer choices: A, B, C, D, or none of the above]

13. Suppose you were interested in how running times tend to change as people get older. Which graph would you use to get an idea of what this trend looks like? [Answer choices: A, B, C, D, or none of the above]

Figure 3. Two sample assessment questions on choosing appropriate statistical display.

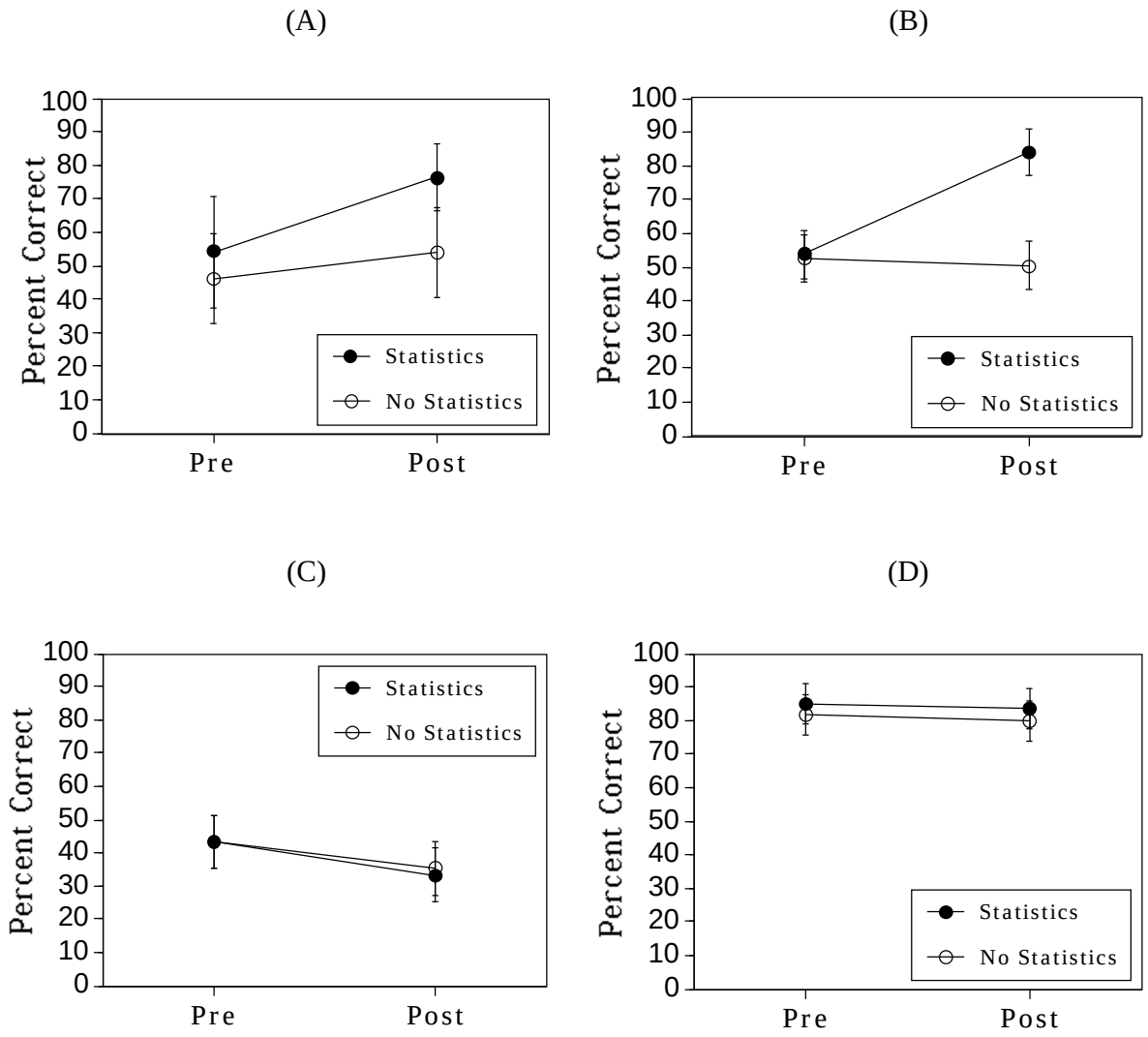


Figure 4. Results of large-scale assessment. Panel A shows percentage correct over all items. Panel B shows percentage correct for items on which students who took the course improved from pre- to posttest. Panel C shows percentage correct for items on which both groups of students showed no improvement, even though they could have. Panel D shows percentage correct for items on which both groups of students showed no improvement, presumably due to a ceiling effect at pretest.

1) Oh, okay. So we need to, he wants to know whether there is a difference in the amount of prize money, the amount of money won by players in the two tours.

2) So, I think this is the prize money, uh, money contains the prize money won by each of these players. Tour indicates which tour the player competes in. Well, you don't really need rank, in order to solve this, right? Cause like, well, I don't know.

4) Um... I'm gonna do a boxplot... ...

5) [Subject uses statistics package to make a boxplot] oh, cool (laugh)- I did it.

6) All right, uh, so just looking at the average. It looks like the people in the senior tour get less money. Um, and there's a lot less variation in the amount of money that, like all the prizes. A couple little outliers in each which means like, I don't know, like some people won, like a lot of money at a time...

Figure 5. Sample protocol with each step coded according to the major steps of statistical reasoning (see Figure 2). Notice that there is no evidence for steps 3 and 7. Also notice that the interpretation is somewhat inaccurate in that boxplots display the median not the mean as a measure of central tendency.

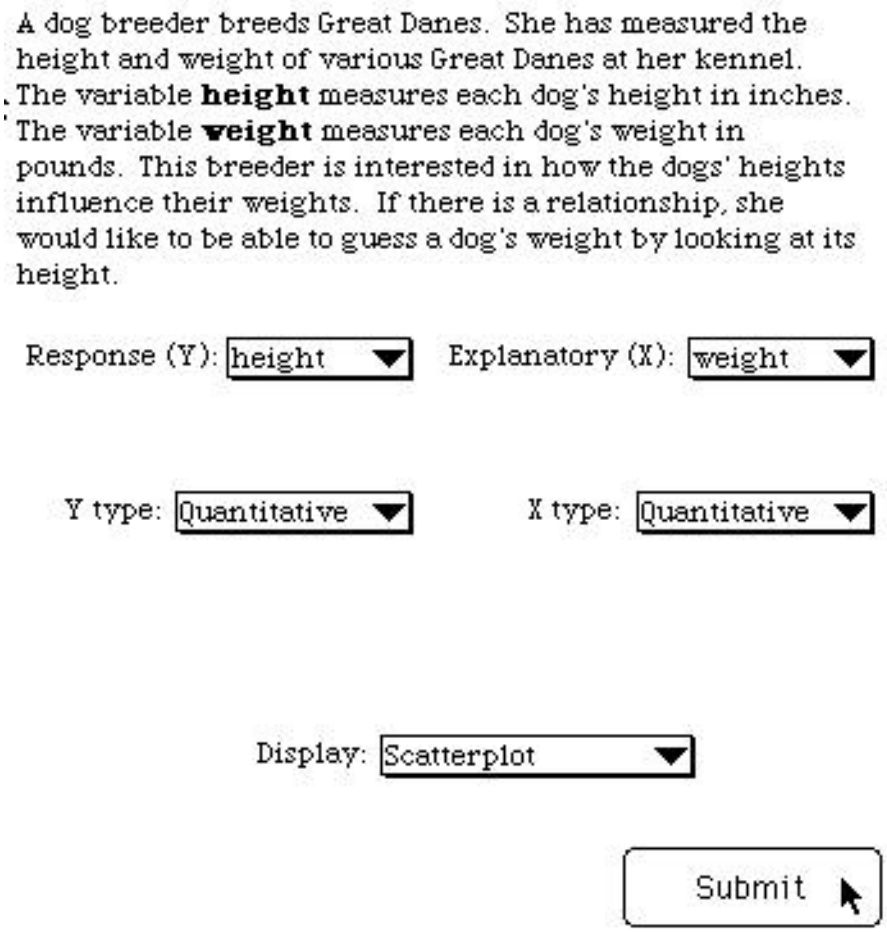
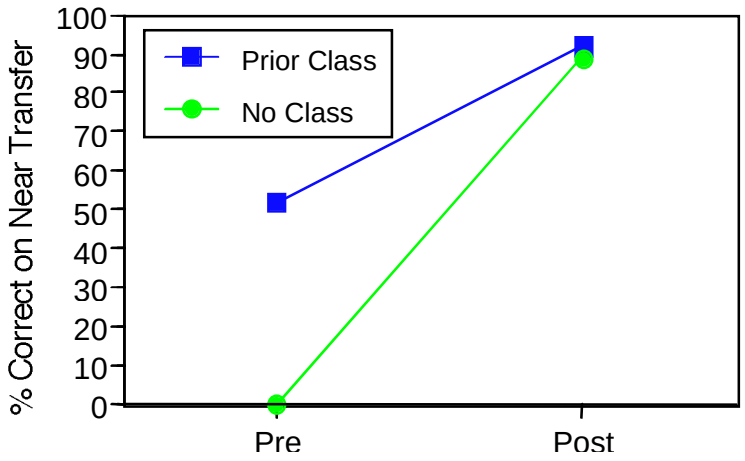


Figure 6. Interface displayed to subjects in learning experiment.

(A)



(B)

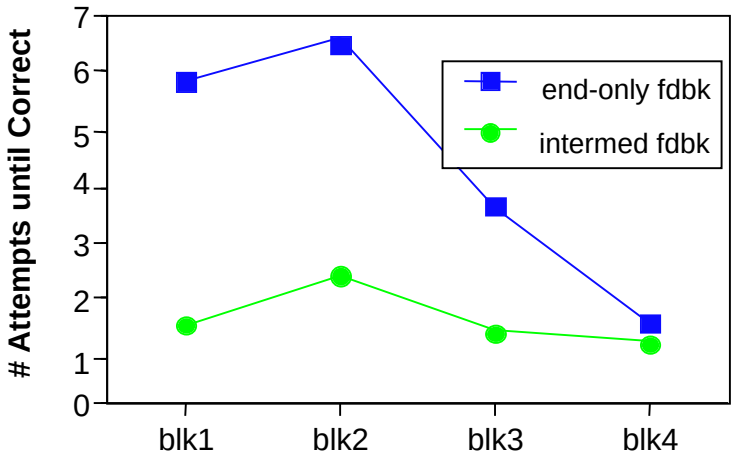


Figure 7. Panel A shows the overall improvement in pre/posttest scores for the two groups of student—those with and without previous statistics courses. Panel B shows the improvement throughout the course of the experiment for the two different conditions (end-only vs. intermediate feedback), including only those students who had no prior statistics classes.

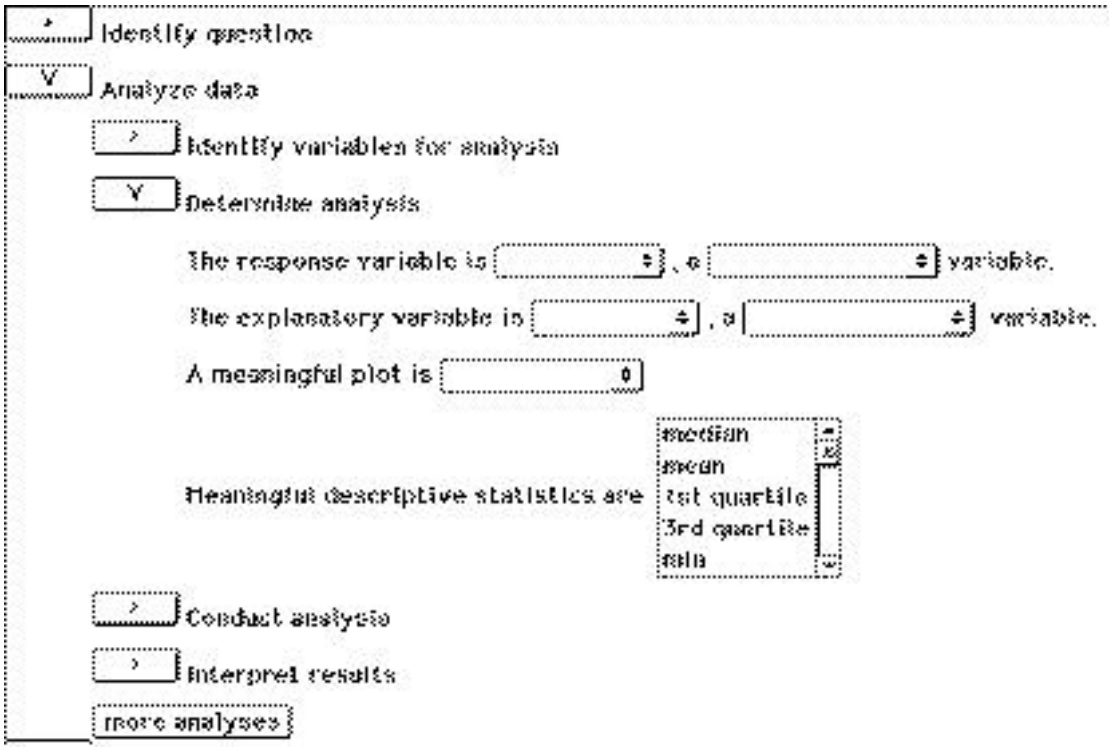


Figure 8. A snapshot of a prototype interface to our learning environment.