EFFECTIVENESS OF SIMULATION TRAINING ON TRANSFER OF STATISTICAL CONCEPTS*

DAVID M. LANE
ZHIHUA TANG
Rice University

ABSTRACT

The effectiveness of simulations for teaching statistical concepts was compared to the effectiveness of a textbook. The variable Medium (simulation versus textbook) and Question specificity (Specific versus Non-specific), were manipulated factorially. Question specificity was defined as follows: Subjects were presented with a scenario in which gumballs were sampled from an urn. Subjects in the "Specific" condition were given a specific question about the outcome of the sampling procedure to consider; subjects in the "Non-specific" condition were asked generally to consider what would happen. A no-treatment control was included. The subjects consisted of 115 college students. The dependent variable was performance on problems requiring subjects to apply what they learned to ill defined everyday problems. Subjects trained by simulation performed significantly better than those trained with a textbook. Subjects in the "Specific" condition performed better than those in the "Non-specific" condition, although the difference did not reach conventional levels of significance. These results support the increasing use of simulation in education and training.

Although simulations have long been known to be effective for training people on tasks with strong perceptual-motor components such as flying a plane [1, 2], their

383

effectiveness in science education is less clear cut [3]. For example, McCloskey and Kohl found that simulations were generally ineffective at improving students' understanding of the path a ball would follow if the string it was attached to broke [4]. Reed [5], on the other hand, found that computer simulations improved students' estimates of answers to algebra word problems, although the effectiveness of simulations was found to depend on other variables such as successive versus simultaneous presentation of examples, active versus passive viewing, and the similarity between examples and test questions. Carlsen and Andre found that learning about electric circuits by simulation led to better conceptual models of the underlying process than learning only by standard text-based instruction [6]. The groups did not differ, however, on the number of correct responses made on a posttest. Rieber and his colleagues have demonstrated the effectiveness of practice on structured simulations in the teaching of physical science [7, 8]. However, Rieber and Parmley found that the effectiveness of a structured simulation illustrating the laws of motion was comparable to the effectiveness of a tutorial but no better [9].

Simulations have played an increasingly important role in statistical education. Cobb noted that teaching laboratories are increasingly incorporating computer simulations illustrating important statistical concepts and laboratory materials that contain statistical simulations are now widely available [10]. For example, ExplorStat [11], Visual Statistics [12], ActivStats [13], HyperStat [14], and the Rice Virtual Lab in Statistics [15] all incorporate simulations for illustrating statistical concepts.

Although there has been little formal evaluation of the effectiveness of simulations for teaching statistics, some positive results have been reported. Wackerly and Lang found students learned better with simulations than did a control group using more traditional materials [11]. DelMas, Garfield, and Chance found a simulation of sampling distributions was an effective teaching tool if used in conjunction with questions focusing the students' attention to relevant features of the simulation [16]. Similarly, de Jong, Hartel, Swaak, and van Joolingen found that students learned better from simulations when given appropriate questions to answer with the simulation [17].

The present study sought to evaluate the effectiveness of a sampling distribution simulation developed by Lane [15]. We were specifically interested in how well students would be able to transfer what they learned about the effects of sample size to real-world situations. Although considerable research on transfer of training has shown that students often fail to apply their knowledge if the superficial aspects of the situation are different from those present during training [18], there is evidence that formal instruction in statistics improves statistical reasoning on everyday statistical problems [19-21]. However, even in these studies, transfer was far from complete, with the mean level of performance on the transfer tasks being far from perfect. We sought to determine whether transfer would be greater for students taught by simulation than by a traditional textbook

approach. Following delMas, Garfield, and Chance [16], we included "question asking" as a variable to see if similar results would occur with our simulation.

Any comparison of the effectiveness of simulation and textbook-based approaches is naturally going to depend, in part, on the quality of the simulation and the quality of the textbook. Therefore, we sought to ensure that the materials used for the textbook-based approach were of very high quality. First, we based the written materials on textbooks that are well regarded and that we judged to be excellent [22-24]. Second, the textbook presentation included examples and graphics appropriate to the concepts being presented. Finally, an audio tape of the material was played as the students followed along in the text so that important statistical concepts and principles could be emphasized vocally as they would be in a live lecture.

METHOD

Subjects

One hundred and fifteen Rice University undergraduate students participated in this experiment in exchange for extra credit in a psychology course. Their ages ranged from seventeen to twenty-two. None of these subjects had ever taken a statistics class.

Subjects were trained and tested in small groups. The assignment of subjects to conditions was determined randomly. This resulted in unequal subject number in each condition but the difference was slight (the maximum difference was 3).

Design

The design consisted of the factorial combination of Medium (computer simulation versus textbook) and Question (Specific versus Non-specific). A no-treatment control was also employed.

Procedures

Upon arrival, the subjects were informed of the general nature of the experiment. The experimenter then administered the 12-minute Wonderlic Personnel Test as a general measure of cognitive ability.

Subjects in the four treatment groups first read two examples and three questions involving each example. All questions were about the relationship between sample size and the resulting sample mean.

One example involved sampling from an urn of thirty-three gumballs, each gumball containing a number from 0 to 32. The second example involved sampling student scores on an achievement test. One of the specific questions for the gumball example was:

> Draw 2 gumballs from the urn and calculate the average of the two numbers on the gumballs. Put the 2 gumballs back to the urn and then draw 10 gumballs out and again calculate the average of their numbers. Which average number is more likely to be bigger than 23?

Subjects were presented with the following three choices: 1) the average of two numbers; 2) the average of ten numbers; and 3) equally probable.

Subjects in the "Non-specific question" condition were presented with the following:

> Draw 2 gumballs from the urn and calculate the average of the two numbers on the gumballs. Put the 2 gumballs back to the urn and then draw 10 gumballs out and again calculate the average of their numbers. What would happen?

In the textbook-based groups, a textbook-style exposition composed of several parts adapted from well-written statistical textbooks [22-24] was presented. Key concepts introduced in the text included sample, sample size, frequency distribution, sampling distribution, and, of particular importance for present purposes, a formal statement of the law of large numbers. In accordance with common practice in statistics training, examples, graphs, formulas, and a table detailing a hypothetical sampling distribution based on the example were included in the text. The example involved the sampling distribution of the mean for achievement test scores. The text was tape-recorded and subjects receiving this training were asked to listen to the tape while viewing the text. This process lasted about twenty-five minutes.

For the computer simulation condition, the training material was centered on an example of random sampling. A script was written and then tape-recorded to ensure that there was no variation across the groups with regard to the presented information and its delivery. The scripts included the key concepts covered in the text version explained in relation to the gumball example. The subjects watched the simulation projected on a big screen in front of them while listening to the tape-recorded script. The computer simulation program was adapted from the Rice Virtual Lab in Statistics, a Web-based statistics training program developed by Lane [15] available at www.ruf.rice.edu/~lane/rvls.html. The simulation contained histogram of the population distribution on the top of the screen. Each time a random sample of a certain size was drawn, the individual scores were depicted on a second histogram, and the sample mean went to a third graph. By repeating this simulating process again and again, the subjects could see the distribution of the sample mean from the third histogram. During training, the experimenter defined two different sample sizes one at a time and the two resulting sampling distributions went to the third and fourth graph respectively. This enabled subjects to see the effect of sample size on the sampling distribution all at once. The whole computer simulation process lasted about thirty minutes.

At the end of the training session, subjects in all the four treatment groups were given the correct answer to the six questions. They were then given a

booklet containing twelve questions and were asked to provide a sensible answer (including an explanation) to each of them. Subjects were also cautioned that the problems may or may not relate to the training they just received. There was no time limit in this posttest session but all the subjects finished the booklet within less than one hour. In general the whole experiment lasted around one and one-half hours for the four treatment groups and one hour for the control group.

Twelve questions adapted from previous research [19, 25] were presented after training. The key dependent variables were the eight questions used by Chen [25] that required the application of the law of large numbers to situations different from the examples used in training. A ninth question used by Chen and considered as near-transfer in that experiment and three filler items were also included. All twelve test items are included in Appendix A. The order of questions was randomized separately for subject with the constraint that Questions 1 and 2 were not presented sequentially, Questions 3 and 4 were not presented sequentially, Questions 5 and 6 were not presented sequentially, and Questions 7 and 8 were not presented sequentially. The reason for this constraint is that we judged the members of these pairs of questions to be somewhat more similar to each other than to members of other pairs.

RESULTS

Coding System

A coding system similar to the one developed by Fong et al. [19] and adopted in other studies [21, 25] was used in this experiment. Every response to the test problems was classified into one of three categories:

1. An entirely deterministic response, that is, one in which the subject made no use of statistical concepts at all. In this category, there was no mention of sample size, randomness, or variance. In addition, an incorrect answer with no explanations was also coded as 1.

Regarding the slot machine question (Appendix A, item 1), a sample answer that falls into this category is: Keith's conclusion is correct. The old man probably owned the machine and that is how he knew which one was better. He just told Keith a lie so Keith would play the machine twice as much as he would have. Keith is correct and the old man lied.

2. A poor statistical response. Response given this score contained some mention of statistical concepts, but were either incomplete or incorrect. A correct answer but with no explanations was also scored as 2.

A sample answer in this category is: You can't conclude anything from just a few tries. The old man is probably right, but the slot machine happened to do the opposite while Keith was there.

3. A good statistical response. This is a response that made the correct use of statistical concepts. Some form of the law of large numbers was used, and the sampling elements were correctly identified.

A sample answer in this category is: Keith's reasoning is flawed—just a few minutes of playing yields very few trials on the slot machines and therefore a greater variance of outcomes. He hasn't determined the actual yield on either machine—only the old man who theoretically has played the machines many times would know. Keith most likely experienced anomalous outcomes.

Two coders (blind with respect to condition) coded a random sample of eighteen (16% of all) booklets separately and they agreed on 83 percent of the answers. This consistency was considered as acceptable and one of the two coders then continued to code the rest of the booklets.

Data from two subjects were discarded because one of them finished only half of the twelve-question set and the other one's handwriting was illegible. Therefore, the subsequent statistical analyses were based on data from the remaining 113 subjects.

Subjects' performance on the eight far-transfer problems were analyzed as Training Condition (5) $\times$ Question (8) analysis of covariance with repeated measures on the last factor.[1] Wonderlic test scores were used as the covariate. There was no evidence of a violation of the assumption of homogeneity of regression slopes, $F(4,103) = 0.75$, $p = 0.558$. Although the questions were of significantly different difficulty, $F(7,756) = 38.46$, $p < 0.01$, there was no evidence of a Condition $\times$ Question interaction, $F(28,756) = 0.87$, $p = 0.656$. The four degrees of freedom in the condition effect were partitioned into: 1) the effect of Medium; 2) the effect of Question specificity; 3) the Medium $\times$ Question specificity interaction; and 4) the difference between the control condition and the average of all four training conditions.

Boxplots of the mean scores averaged across the eight problems (adjusted for the covariate) as a function of condition are shown in Figure 1. As can be seen in Figure 1, the control group performed worse than the training groups; the difference between the control condition and the average of the training conditions was significant, $F(1,107) = 21.98$, $p < .01$. Dunnett's test was used to test the difference between each of the four experimental groups and the control group. Three of the four experimental conditions were significantly higher than the control condition ($p < .05$), the only exception being the textbook/non-specific question. These results are consistent with those of Nisbett and others and

---

[1]The ninth question, considered near transfer in the Chen [25] study, is actually a far transfer problem in this study. We followed our *a priori* plan was to make the first eight questions the dependent variables. It should be noted, however, that inclusion of the ninth question makes no substantive difference in the analysis.

provide further evidence that formal statistical training increases subjects' statistical reasoning in ill-defined everyday problems.

Of primary concern in the present study was the effect of simulation training. Figure 1 shows the simulation condition produced better performance than did the textbook condition. This difference was significant, $F(1,107) = 5.84$, $p = .017$. Subjects trained with the simulation outperformed subjects trained with the textbook on seven out of the eight critical questions. Although performance was higher in the specific-question condition than in the non-specific question condition, the effect only approached significance, $F(1,107) = 3.59$, $p = .061$. The Medium × Question Specificity interaction did not approach significance, $F(1,107) = 1.86$, $p = .176$, although there was some hint that the simulation's advantage is greater for non-specific questions.

DISCUSSION

Formal statistical training transferred to problems of every-day reasoning. This finding is inconsistent with the formerly-prevalent view that people use only domain-specific rules in solving concrete problems [26], and consistent with studies showing that people can learn abstract rules and apply them to ill-defined everyday problems [19, 21].

This study's most important finding is that training by simulation led to better performance than training using a traditional textbook approach. Subjects trained with the simulation were apparently more able to recognize the key elements of ill-defined problems embedded in various real world situations and apply the relevant statistical principles. This provides support for the thesis that simulation is effective for training on educational and cognitive tasks (as opposed to tasks such as flying an airplane where simulation has been shown unequivocally to be effective).
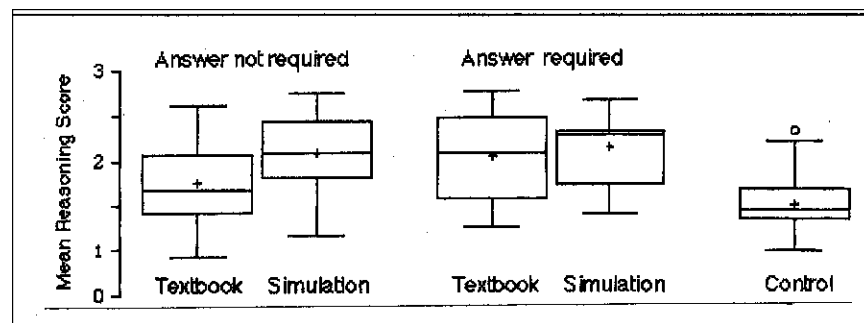

Figure 1. Mean reasoning score as a function of
experimental condition.

There was a strong hint that asking subjects a specific question to consider during training was advantageous, particularly in the textbook condition, although the effect was not statistically significant at conventional levels. Previous research has found that giving the subjects a question to focus on during training is effective in otherwise unstructured discovery learning [3]. The effect of providing a question may have been reduced in the present study because all training regimens were structured and all subjects were given the same explanations of the answers to the questions immediately after training.

Despite the significant effect of training, many subjects had at least some difficulty in applying the statistical concepts and principles that they had just learned even in the most effective training conditions. The mean score on the eight critical questions was not much greater than two on a three-point scale even in the "Simulation/Specific question" condition. This supports Detterman's conclusion that far transfer is very difficult to produce [18].

An interesting aspect of the current research is in that a simulation lasting only thirty minutes and consisting of no more than a purely graphical demonstration of the sampling procedure led to better performance than the textbook condition that included the same kind of information. A possible explanation of the superior performance of the simulation-based groups may rely on the characteristics of the training content in the current study. Statistics has a long-standing reputation of being abstract and thus hard to understand for naïve students, and the concept of a sampling distribution and its relationship to the law of large numbers is one of the more abstract concepts encountered by a beginning student. However, these concepts can be understood concretely by means of an animated simulation demonstrating the sampling process. Therefore, the effectiveness of computer simulation found in the current study may be partly attributable to the compatibility between the delivery medium and the essence of the training content.

A second possibility is that the computer simulation emphasized certain aspects of the problems covered in this study. In addition to helping subjects gain an insight of the statistical property of the law of large numbers, the concrete process of drawing samples out of the population and graphically depicting the outcome of each sample mean made the key element of the problem, namely the randomness of these events, even clearer. This is exactly the first step that the subjects need to recognize in order to successfully solve the post-training real-world problems.

A third factor that might contribute to the superior performance of the simulation groups is motivational. Non-conventional teaching methods such as simulations can enhance students' performance because they are more interesting and thus more engaging than pure textbooks. Although in this study we didn't measure subjects' motivation, it is our observation that subjects in the simulation groups appeared to be more engaged throughout the training session.

Although the simulation was shown to be effective in a test taken immediately after training, it is not certain that the effect would persist over time. As some

researchers argued, acquisition performance is at most an imperfect indicator of learning, and the effect of any training program can not be fully tapped without considering both transfer and retention [27]. Learning conditions resulting in immediate superior performance are not necessarily those that lead to better retention over time [28, 29].

Another aspect that warrants investigation is the computer simulation technique itself. In the current study the simulation served exclusively as a demonstration of the information that was otherwise conveyed through text. The subjects in the simulation groups viewed the pre-defined simulation procedure without being able to explore or interact with it on their own. Some researchers consider relevant and sustained user interactivity the very essence of a simulation program [3, 9, 30, 31] and argue that the possibility for a learner to infer, through experimentation, characteristics of the model underlying the simulation is critically important. Therefore, it is likely that the advantage of simulation in the present experiment would have been greater if learners had had the opportunity to interact with the simulation. Further research is indicated to test this hypothesis.

Taken together, the current results demonstrate clearly that, compared to a textbook approach, computer simulation can be a very effective way in teaching college students important statistical concepts and principles. However, this result should be applied with caution. The fast development of information technology has opened up many opportunities for instructors to adopt new instructional techniques. Meanwhile we believe that the traditional way of classroom instruction is far more sophisticated than just a textbook approach. An experienced instructor would employ every technique possible to ensure the quality of student learning, such as asking questions, giving demonstrations, providing hands-on practices, etc. Therefore, the task that practitioners are facing with is how to effectively integrate new technology into the traditional classroom setting. Before we can do that, however, more empirical research efforts are needed to fully address the question of potential benefits of these new techniques.

APPENDIX A
Questions Used in Post-Training Test

Questions 1 through 8 are the critical items that demand far-transfer of what the subjects had learned about the law of large numbers. Question 9 is the near-transfer item from Chen [25]. Questions 10 through 12 are filler items.

1. Keith was driving through Nevada when he stopped in a gas station with two slot machines. An old man nearby said, "There ain't no winning system for slot machines. It's all luck. You just put in a coin, pull the lever and hope you win. But let me tell you this: some machines are easier to lose on than others are. That's because the owners can change the mechanism of the slots so that some of them will be more likely to make you lose. See those two machines? The one on the left will give you even chances of winning, but the one on the right is fixed to make you

lose much more often than you win. Take it from me. I've played them for years."
Keith played both machines for a few minutes. On the left machine, he lost twice as
much as he won. On the right machine, he won twice as much as he lost. He
concluded that the old man was wrong about the odds of winning on the two slot
machines; the opposite was true, the one on the right was more favorable than
the one on the left.

Assume the old man has nothing to gain from the slot machines. Comment on
Keith's conclusion and reasoning. Do you agree? Explain your answer.

2. There are two hospitals in Normaltown, TX: Central and Valley. At Central
Hospital, 100 babies are born a day. At Valley Hospital, 30 babies are born a day.

Are days with more than 60% female births:
1) More common at Central hospital;
2) More common at Valley hospital;
3) Equally common at either hospital.
Explain your position.

3. The Caldwells were looking for a "safety-conscious, Swedish" car. As luck
would have it, their old car stopped working on the last day of the closeout sale
for the model year for both the Volvo and Saab. They quickly got out their
Consumer Reports where they found that the consensus was that both cars were
very sound mechanically, although the Volvo was felt to be slightly superior on
some dimensions. They also found that readers of Consumer Reports who owned
Volvo's reported having somewhat fewer mechanical problems than owners of
Saabs. They were about to make a deal when Mr. Caldwell remembered that they
had two friends who owned a Saab and one who owned a Volvo. Mr. Caldwell
called up the friends. Both Saab owners reported having a few mechanical
problems, but nothing major. The Volvo owner exploded when asked how he liked
his car. "First that fancy fuel injector computer thing went out: $250 bucks. Next I
started having trouble with the rear end. Had to replace it. Then the transmission
and the clutch. I finally sold it after 3 years for junk."

Given that the Caldwells are going to buy either a Volvo or Saab today, which
do you think they should buy and why?

4. The psychology department keeps records on the performance of thousands
of its graduate students and relates this performance to scores on all kinds of
background information about the students. Recently, there was a student from a
small college with GRE scores and a GPA such that almost all accepted students
had scores as high or higher, while rejected students typically had lower scores.
The letters of recommendation were quite good, but none of the writers of the
letters were personally known to the reviewers. One of the reviewers argued
against admission because students from small colleges tend to perform below the
department average. Another disagreed, noting that one student admitted years

ago from a small college also had similar scores, but became one of the top three students in the department.

Comment on the arguments put forth by the two reviewers. What are their strengths and weaknesses? Should the student be admitted and why?

5. The director of a Broadway play just finished auditions for the female lead. Two of the candidates gave readings for the part that were very good. The third was given by an actress he had worked with in four previous plays. The director thought she had been superb in each. Unfortunately, of her four readings, one was very good, but the other three had been flat.

The director needs to choose the female lead immediately. What should the director do? Hire the third actress or hire one of the two whose readings he liked better? Why?

6. Coach Graves has been scouting two different quarterbacks for his college team, Mike and Scott. According to high-school records, Mike has performed excellently and seems to have a promising future. Scott's records show he has been a little less productive. Although he does show some promise, his history suggests that he will become a back-up quarterback. During the two-day try-outs, Mike has difficulty completing passes, while Scott sails through easily. His assistants tell Coach Graves that Mike is performing poorly, while Scott is doing very well.

Coach Graves needs to choose the best quarterback he can get and he can only choose one. Which quarterback should Coach Graves choose and why?

7. At a bar, Mark and Joe are about to play one tie-breaker game of darts. Both know that Mark is better with darts, so they agree that Joe can choose the rules for the last game. Joe has to play by the same rules he chooses for Mark.

How can Joe maximize his odds of winning? Explain your reasoning.

8. In the last five years, the stock market of a certain country has been in decline, with investors losing an average of 5% each year. Assuming that Person S had chosen to invest his money in a single company and Person D had chosen to invest his money in ten different companies, compare the performance of the two investors for the last five years.

How did the two investors perform?

9. Shady Jake, a well-known gambler, was playing a game with three dice when he announced that the house was using loaded dice. Jake claimed that the dice had been averaging 12 per roll for the last five hours when they should be averaging 10.5 per roll. The pit boss claimed that the dice were fair, pointing out that the last four rolls had averaged 10.5.

If both players are telling the truth about the averages, who should the authorities believe and why?

10. An economist was arguing in favor of a guaranteed minimum income for everyone. He cited a recent study of 700 people in the Untied States with inherited wealth. Nearly 92% of those people, he said, worked at some job that provided earned income sufficient to provide at least a middle-class life style. The study showed, he said, that contrary to public opinion, people will work in preference to being idle. Thus a guaranteed income policy would result in little or no increase in the number of people unwilling to work.

Comment on the economist's reasoning.

11. A brewery buys nearly all of its glass bottles from a local manufacturer. One summer, the local company is unable to deliver enough bottles and the brewery orders a shipment from a large glass manufacturer that distributes its products nation-wide. On the first day that these new bottles are used, however, the bottle-filling machinery has to be stopped four times because of jamming, and as a result, production for the day is unusually low. (Ordinarily, the brewery does not experience more than one jamming a day, and frequently there are none at all.) The foreman is worried about the new bottles. He decides to test the new bottles produced by the national manufacturer. He randomly selects 300 cases of these new bottles and instructs the bottle-filler operators to record carefully each jamming incident. Meanwhile, company mechanics carefully lubricate and check adjustments on the bottle-filling machinery. When they are finished the machinery is running more smoothly than it has for years. During the next 2 days, the 300 cases of new bottles are fed to the machine. There are only two jamming incidents, one each day. The foreman concludes that there is little or no real disadvantage of the new bottles with respect to jamming of the bottle-filling machinery.

Comment on the foreman's reasoning. Is it basically sound? Can his procedure be criticized?

12. Martha was talking to a fellow passenger on an airplane. The fellow passenger was on his way to Hawaii for a month's vacation. "I don't like vacations myself," Martha said. "I've always worked. I put myself through college and law school and now I have a full-time legal practice. Frequently, of course, I've had slow periods when I wasn't working at all, but I never liked those times. For example, there would usually be a week or two between the end of school and the beginning of a summer job and another week or two of enforced idleness at the end of the summer. And there were many occasions when I was getting started in my career when I had no real work to do for fairly long periods. But I never enjoyed the leisure. I know there are some people who talk about using vacations to 'recharge' themselves, but I suspect many of these people don't really enjoy their work or don't have a very high energy level. I do have a lot of energy, and I do enjoy my work, and I guess that's why I don't really like vacations."

Analyze Martha's reasoning. Do you think she had good evidence for feeling she doesn't like vacations?

REFERENCES

1. R. E. Flaxman and E. A. Stark, Training Simulators, in *Handbook of Human Factors,* G. Salvendy (ed.), John Wiley & Sons, Inc., pp. 1039-1058, 1987.
2. H. H. Valverde, A Review of Flight Simulator Transfer of Training Studies, *Human Factors, 15,* pp. 510-523, 1973.
3. T. de Jong and R. van Joolingen, Scientific Discovery learning with Computer Simulations of Conceptual Domains, *Review of Educational Research, 68,* pp. 179-201, 1998.
4. M. McCloskey and D. Kohl, Naive Physics: The Curvilinear Impetus Principle and Its Role in Interactions with Moving Objects, *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* pp. 146-156, 1983.
5. S. K. Reed, Effect of Computer Graphics on Improving Estimates to Algebra Word Problems, *Journal of Educational Psychology, 77,* pp. 285-298, 1985.
6. D. D. Carlsen and T. Andre, Use of a Microcomputer Simulation and Conceptual Change Text to Overcome Student Preconceptions about Electric Circuits, *Journal of Computer-Based Instruction, 19*:4, pp. 105-109, 1992.
7. L. P. Rieber, Using Computer Animated Graphics in Science Instruction with Children, *Journal of Educational Psychology, 82,* pp. 135-140, 1990.
8. L. P. Rieber, M. Boyce, and C. Assad, The Effects of Computer Animation on Adult Learning and Retrieval Tasks, *Journal of Computer-Based Instruction, 17,* pp. 46-52, 1990.
9. L. P. Rieber and M. W. Parmley, To Teach or Not to Teach? Comparing the Use of Computer-Based Simulations in Deductive versus Inductive Approaches to Learning with Adults in Science, *Journal of Educational Computing Research, 13,* pp. 359-374, 1995.
10. G. Cobb, Reconsidering Statistics Education: A National Science Foundation Conference, *Journal of Statistics Education, 1,* 1993.
11. D. Wackerly and J. Lang, *ExplorStat,* Paper presented at the Joint Statistics Meeting, Chicago, August 1996.
12. D. P. Doane, K. Mathieson, and R. L. Tracy, *Visual Statistics,* McGraw-Hill, Boston, Massachusetts, 1997.
13. P. Velleman, *ActivStats,* Addison Wesley Longman, Reading, Massachusetts, 1998.
14. D. M. Lane, *HyperStat: Hypermedia for Learning Statistics and Analyzing Data,* Academic Press, New York, 1993.
15. D. M. Lane, The Rice Virtual Lab in Statistics, *Behavior Research Methods, Instruments, and Computers, 31,* pp. 24-33, 1999.
16. R. delMas, J. Garfield, and B. Chance, Assessing the Effects of a Computer Microworld on Statistical Reasoning, *Proceedings of Fifth Conference of the International Council of Teachers of Statistics* (ICOTS5), Singapore, July 1998.
17. T. de Jong, H. Hartel, J. Swaak, and W. van Joolingen, Support for Simulation-Based Learning: The Effects of Assignments in Learning about Transmission Lines, in *Computer Aided Learning and Instruction in Science and Engineering,* A. Diaz de Ilarazza Sanchez, and I. Fernandez de Castro (eds.), Springer-Verlag, Berlin, Germany, pp. 9-27, 1998.
18. D. Detterman, The Case for the Prosecution: Transfer as in Epiphenomenon, in *Transfer on Trial: Intelligence, Cognition, and Instruction,* D. Detterman and R. Sternberg (eds.), Ablex Publishing Company, Norwood, New Jersey, pp. 1-24, 1993.

19. G. T. Fong, D. H. Krantz, and R. E. Nisbett, The Effects of Statistical Training on Thinking about Everyday Problems, *Cognitive Psychology, 18,* pp. 253-292, 1986.

20. G. T. Fong and R. E. Nisbett, Immediate and Delayed Transfer of Training Effects in Statistical Reasoning, *Journal of Experimental Psychology: General, 120,* pp. 34-45, 1991.

21. P. Kosonen and P. H. Winne, Effects of Teaching Statistical Laws on Reasoning about Everyday Problems, *Journal of Educational Psychology, 87,* pp. 33-46, 1995.

22. R. Lowry, *The Architecture of Chance: An Introduction to the Logic and Arithmetic of Probability,* Oxford University Press, New York, 1989.

23. K. M. Rosenberg, *Statistics for Behavioral Sciences,* Wm. C. Brown, Dubuque, 1990.

24. C. Spatz, *Basic Statistics: Tales of Distributions,* Brooks/Cole Publishing Company, Pacific Grove, California, 1993.

25. B. Chen, *Transfer Between Different Contexts: Examining the Effect of Interaction,* unpublished Master's thesis, Rice University, 1999.

26. R. A. Griggs and J. R. Cox, The Elusive Thematic-Materials Effect in Wason's Selection Task, *British Journal of Psychology, 73,* pp. 407-420, 1982.

27. R. A. Schmidt and R. A. Bjork, New Conceptualizations of Practice: Common Principles in Three Paradigms Suggest New Concepts for Training, *Psychological Science, 3,* pp. 207-217, 1992.

28. J. B. Shea and R. L. Morgan, Contextual Interference Effects on the Acquisition, Retention, and Transfer of a Motor Skill, *Journal of Experimental Psychology: Human Learning and Memory, 5,* pp. 179-187, 1979.

29. T. K. Landauer and R. A. Bjork, Optimum Rehearsal Patterns and Name Learning, in *Practical Aspects of Memory,* M. M. Gruneberg, P. E. Morris, and R. N. Sykes (eds.), Academic Press, London, pp. 625-632, 1979.

30. P. W. Grimes and T. E. Willey, The Effectiveness of Microcomputer Simulations in the Principles of Economics Course, *Computers & Education, 14,* pp. 81-86, 1990.

31. R. H. Rivers and E. Vockell, Computer Simulations to Stimulate Scientific Problem Solving, *Journal of Research in Science Teaching, 24,* pp. 403-415, 1987.