RICE UNIVERSITY

Learner Control in an Interactive Learning Environment

by

Zhihua Tang

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

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

David M. Lane, Chair
Associate Professor of Psychology, Statistics and Management

Michelle Hebl
Assistant Professor of Psychology and Management

H. Albert Napier
Professor of Management and Psychology

Miguel A. Quinones
Associate Professor of Psychology and Management

Houston, Texas

December, 2003
ABSTRACT

Learner Control in an Interactive Learning Environment

by

Zhihua Tang

The emergence of computer-based interactive learning environments has presented us with many unanswered questions. The current research examined learner control in an interactive learning environment from two perspectives.

In Part I, three experiments were conducted to compare simulation-based interactive learning with expository learning in learning statistics. In Experiment 1, interactive learning was compared to textbook-based expository learning. Interactive learning was structured in two different ways so that learners received either directive or nondirective guidance while interacting with the simulation. Compared to expository learning, learner control resulted in slightly improved but much more consistent performance on a knowledge test as well as more positive affect towards learning. In Experiment 2, learner control was compared to simulation-based expository learning. In each learning condition, half of the participants were further asked to predict simulation outcomes during the learning process. Interactive learning resulted in significantly higher response accuracy on the knowledge test than did expository learning. It also improved learners' performance on a transfer test for those with medium lower cognitive ability. Making predictions was more beneficial for interactive learning than for expository learning. Experiment 3 examined the effects of interactive learning over time. The
expository learning group was yoked with the interactive learning group by passively observing their interaction with the simulation. Participants were tested either immediately after learning or after a one-week delay. Performances of the interactive learning group remained stable over this period of time. However, learner control did not improve learners' performance compared to expository learning. Reasons for this finding were discussed.

In Part II, two iterations of user testing were conducted to examine user interaction with the Connexions Web-based learning environment. User interaction was considered an integral part of learner control in such a complex environment. Usability information gathered from user testing was used to aid the software development effort. The current research supported the idea that learner control can lead to better learning than expository learning but emphasized the importance of learning structure and potential aptitude-treatment interaction in simulation-based interactive learning. These findings have implications for larger-scale interactive learning environments, such as Web-based learning, as well.
ACKNOWLEDGMENTS

Funding for this project was graciously provided by the CLASS Foundation through a Doctoral Dissertation Award. I would like to thank the Connexions project for giving me the opportunity to work on this project, and as a result, to extend my research. Many thanks go to Heidi Ziemer, who kindly provided convenience for conducting the experiments. I wish to thank my committee for their helpful comments. I am very grateful to my advisor, David Lane, for his support and guidance. I would also like to express my gratitude to my parents, for life and everything. Finally, I thank the Lord for his grace and blessing.
TABLE OF CONTENTS

Background: A Brief Review of Three Learning Paradigms ................................................. 1

Part I: Learning Statistics with an Interactive Computer Simulation ................................. 6

1. Introduction ......................................................................................................................... 6

2. Experiment 1 ...................................................................................................................... 15

   2.1 Method .......................................................................................................................... 15

      2.1.1 Participants ............................................................................................................. 15

      2.1.2 Materials ............................................................................................................... 16

      2.1.3 Procedure .............................................................................................................. 19

      2.1.4 Data Analysis ......................................................................................................... 20

      2.1.5 Transfer Question Coding Scheme ....................................................................... 22

   2.2 Results .......................................................................................................................... 23

   2.3 Discussion ..................................................................................................................... 28

      2.3.1 Aptitude-Treatment Interaction in Interactive Learning ..................................... 32

3. Experiment 2 ...................................................................................................................... 37

   3.1 Method .......................................................................................................................... 39

      3.1.1 Participants ............................................................................................................. 39

      3.1.2 Materials ............................................................................................................... 39

      3.1.3 Procedure .............................................................................................................. 41

      3.1.4 Data Analysis ......................................................................................................... 42

   3.2 Results .......................................................................................................................... 43

   3.3 Discussion ..................................................................................................................... 49
4. Experiment 3 ................................................................. 53
  4.1 Method ................................................................. 54
    4.1.1 Participants .................................................... 54
    4.1.2 Materials ....................................................... 54
    4.1.3 Procedure ..................................................... 55
    4.1.4 Data Analysis ................................................ 56
  4.2 Results .............................................................. 57
  4.3 Discussion .......................................................... 60

5. General Discussion .................................................. 65

Part II: Evaluating the Usability of a Web-Based Learning Environment .............................................. 73
  1. Introduction .......................................................... 73
  2. User Testing 1 ........................................................ 78
    2.1 Method ............................................................. 78
      2.1.1 Participants .................................................. 78
      2.1.2 Materials .................................................... 78
      2.1.3 Procedure ................................................... 80
      2.1.4 Data Analysis ............................................... 80
    2.2 Results and Discussion ....................................... 80
  3. User Testing 2 ........................................................ 83
    3.1 Method ............................................................ 83
      3.1.1 Participants .................................................. 83
      3.1.2 Materials .................................................... 83
3.1.3 Procedure and Data Analysis ................................. 84
3.2 Results and discussion ........................................ 85

4. General Discussion .............................................. 88

Conclusions ......................................................... 92

References .......................................................... 94

Appendix A: Excerpts of Directive and Nondirective Guidance for Interactive Learning in Experiment 1 ........................................ 98

Appendix B: Multiple-Choice Questions Used in Experiments 1, 2 and 3 ......................... 99

Appendix C: Scenarios Used in the Transfer Test in Experiments 1, 2 and 3 ................. 101

Appendix D: Questionnaire Used in Experiments 1, 2 and 3 ................................. 105

Appendix E: Sample Worksheet Provided in the Interactive Learning Conditions in Experiments 2 and 3......................................... 106

Appendix F: ANCOVA Results from Experiment 2 ........................................ 107

Appendix G: Tasks Used in User Testing 1 and 2........................................ 108
LIST OF TABLES

Table 1. Average ratings on subjective measures in the three learning conditions in Experiment 1 ................................................................. 27

Table 2. Average ratings on subjective measures in the four learning conditions in Experiment 2 ................................................................. 49

Table 3. Average ratings on subjective measures in the four experimental conditions in Experiment 3 ................................................................. 60

Table 4. Observation of user performances in user testing 1 ......................................................... 81

Table 5. Observation of user performances in user testing 2 ......................................................... 85

Table 6. Comparison of the most difficult tasks in the two rounds of user testing .......... 86
LIST OF FIGURES

Figure 1. User interface of the Sampling Distribution simulation ........................................... 17

Figure 2. Frequency of interacting with the simulation as a function of learning structure in Experiment 1.................................................................................. 23

Figure 3. Accuracy on the statistical knowledge test in the three learning conditions in Experiment 1 ........................................................................................................... 24

Figure 4. Statistical reasoning as a function of experimental condition in Experiment 1 ......................................................................................................................... 26

Figure 5. Accuracy on the statistical knowledge test in the four treatment conditions in Experiment 2 ........................................................................................................... 44

Figure 6. Accuracy on the knowledge test as a function of cognitive ability and learner control in Experiment 2 ................................................................................. 45

Figure 7. Statistical reasoning as a function of experimental condition in Experiment 2 ......................................................................................................................... 46

Figure 8. Statistical reasoning as a function of learner control and cognitive ability in Experiment 2 ........................................................................................................... 47

Figure 9. Accuracy on the statistical knowledge test as a function of experimental condition in Experiment 3 ......................................................................................... 58

Figure 11. Screenshot of the initial Connexions student user interface ...................................... 79

Figure 12. The improved student user interface based on results from the first user testing .................................................................................................................... 84

Figure 13. The improved student user interface based on results from the second user testing .................................................................................................................. 88
BACKGROUND: A BRIEF REVIEW OF THREE LEARNING PARADIGMS

The design and evaluation of a learning environment is based on principles of learning and instruction. Each learning environment reflects our basic understanding of how individuals think and learn. As computer-based instruction is more and more driven by the advances in technology, it has become increasingly important to review the psychology of learning in order to fully appreciate the values of different types of learning environment.

Historically, three schools of psychology have had significant influence on how learning is defined, studied, and understood. These are behavioral psychology, cognitive psychology, and constructivist psychology. Each school has its own prescription for what would be expected to be the most effective learning environment.

Behavioral psychology

Behavioral psychology restricts itself to the study of observable behaviors and environmental events, to the exclusion of non-observable constructs such as memory, beliefs, or mind. It defines learning as relatively permanent behavioral changes associated with the contiguous linking of stimuli, responses, and reinforcers (Skinner, 1957). In the behavioral paradigm, learning is viewed as the outcome of practice and reinforcement, and is influenced by external motivation. Regarding pedagogy, behavioral psychology assumes that observations, listening to explanation, or engaging in activities with feedback will result in learning. Consequently, learner is seen as a passive participant in
the teaching-learning process while much of the responsibility falls on the instructor, who puts great efforts into organizing the curriculum and explaining the content to the learner.

**Cognitive psychology**

In contrast to behaviorism, cognitive psychology assumes the existence of unobservable constructs such as mind, thinking, memory, and motivation, and emphasizes their roles in learning. In the cognitive paradigm, the learning process is seen as one of information processing, which consists of intervening stages of selecting, encoding, organizing, storing, retrieving, decoding, and generating information. Consequently, learning is under the influence of a multitude of factors.

The most important difference between the cognitive and behavioral paradigms is that, in the cognitive approach, learning is seen as an active rather than a passive process. People learn not only by observing and listening to explanations, but also by actively interacting with the material.

In addition to this tenet of active learning, cognitive psychology also emphasizes a number of other factors as important for learning, including motivation, transfer, and individual differences. From the cognitive perspective, a learner's motivation is induced not just by extrinsic motivators but more importantly by elements intrinsic to the learning process. Techniques for enhancing intrinsic motivation include arousing learner's curiosity, presenting proper challenge, and giving learner control over the learning process (e.g., Malone & Lepper, 1987). Transfer of learning refers to one's ability to apply what is learned in an educational setting to real-world situations. Transfer can be enhanced by building variation into the learning environment so as to facilitate
generalization. Individual differences are stressed in the cognitive approach because not all people learn alike or make progress at the same rate. Learners differ from each other on many characteristics such as need, motivation, cognitive ability, learning style, and so forth. These differences result in individualized learning experiences and affect the effectiveness of a learning program.

Constructivist psychology

Constructivism is the opposite of objectivism. Objectivism holds that there is an objective world that one perceives more or less accurately through our senses, and learning is the process of interpreting information coming through the senses and responding to events existing in the objective world. In contrast, constructivism contends that people add their individual interpretation to what they perceive; therefore, knowledge is constructed by each individual instead of coming from an independent outside existence.

Constructivist psychology emphasizes the learner’s role in the learning process to a much greater extent than the cognitive approach. It defines learning as a process of an individual actively constructing knowledge. The learner acquires an idea through his/her own insight. Therefore, learning is the creation of its own subject matter rather than a simple acquisition of what was already there. Constructivism stresses that in order to facilitate knowledge construction, the learner should be given opportunity to explore, experiment, and research in the learning environment.

Constructivist psychology supports the idea of discovery learning, which was promoted by the work of Bruner (1961). However, in practice pure discovery learning
often posed serious problems for the majority of learners and rarely achieved any meaningful success. Therefore, the current constructivist approach emphasizes guided discovery learning. The learning environments are structured in certain ways so that learners receive various forms of learning support during their exploration.

The psychology of learning has evolved considerably over time. The development of computer-based learning environment reflects the changes in learning paradigm. The earliest computer applications in the educational settings, such as tutorials and drills, were designed to meet the behaviorist goal of learning. By contrast, more recent educational technology has embraced the idea of active learning by putting greater emphasis on learner interactivity. The most recent development, such as simulations, educational games, hypermedia and Web-based learning, all allow for much increased learner control.

The emergence of interactive learning environments has been, to some extent, driven by the rapid advances in information technology. However, interactive learning environments are less well researched and understood than those on the lower end of the interactivity continuum, such as drills and tutorials. Therefore, there is a great need to evaluate, revise and fine-tune these interactive learning environments in order to ensure their effectiveness.

The current research aimed to investigate learner control in an interactive learning environment from two perspectives. In Part I, three experiments were conducted to investigate the constraints and effectiveness of learner control in the context of learning statistics with a computer simulation. In Part II, user interaction in a Web-based learning
environment was examined through two iterations of user testing, and the usability information gathered was used to aid the iterative design of the system.
PART I: LEARNING STATISTICS IN
A SIMULATION-BASED LEARNING ENVIRONMENT

1. INTRODUCTION

A computer simulation used for science education is a software program that contains a conceptual model of a system. The conceptual model holds concepts, facts, and principles related to the system being simulated. In a simulation-based learning environment, a student learns the conceptual model by repeatedly changing values of the input variables, observing the corresponding changes in values of the output variables, and making inference about the characteristics of the underlying model (de Jong & van Joolingen, 1998).

Both the cognitive and constructivist learning paradigms support the use of simulation-based learning environments in science education because, by interacting with the computer program, the learner becomes an active participant in the learning process. Particularly, simulation-based environments are considered by researchers with a constructivist view as being well suited for scientific discovery learning. De Jong and van Joolingen (1998) drew an analogy between simulation-based learning and scientific research. They observed that, in a simulation-based environment, learners are essentially engaged in simulated research by applying scientific methods, conducting repeated experiments, and eventually reaching an understanding of the underlying principle of the system. Their activities closely follow the steps involved in scientific reasoning, including generating hypothesis, designing experiments, interpreting data, and at a higher level, regulating the discovery learning process use systematic planning and monitoring.
By encouraging learners active participation, simulation-based learning has two distinctive advantages over passive learning methods such as those based primarily on books, lectures, and tutorials. First, simulation enhances learner's motivation. Malone and Lepper (1987) maintained that intrinsic motivators that are inherent to the process of learning are more beneficial than extrinsic motivators as stressed in the behavioral paradigm. They proposed four principles to enhance learner's intrinsic motivation: (a) present challenge at a level that suits the individual; (b) arouse curiosity; (c) give the learner control over the learning process; and (d) satisfy the learner's fantasies. The first three elements are well maintained in most simulations designed for scientific education. The second advantage of simulation is its potential to facilitate transfer of learning. Research has shown that practice variability enhances transfer. A computer-based simulation accommodates and encourages various learner input. As the learner experiences more variability, he/she is more likely to achieve a deeper understanding of the system and further to generalize learning to other contexts (Alessi & Trollip, 2001).

Empirical research on simulation-based interactive learning has been conducted following a number of paths. Early research efforts in this area focused on comparing learning from simulation with learning from certain forms of expository learning, such as tutorials or classroom lecture. In some studies, a simulation was used as the sole content delivery medium and was compared to expository learning (e.g., Grimes & Willey, 1990). In others, a simulation was embedded in a curriculum and combined with expository learning, and the learning outcome was compared to that of the original curriculum (e.g., Rivers & Vockell, 1987). Research on simulation-based learning has been conducted in a range of subject domains, including biology, physics, and
economics. While results from some research favored simulation-based learning (e.g., Grimes & Willey, 1990), other studies failed to find differences between simulation-based and expository learning (e.g., Carlsen & Andre, 1992). The overall results from empirical research indicated that there was no clear-cut answer regarding the effectiveness of simulation-based interactive learning.

In many of the studies on simulation-based learning, researchers looked at the learning process and noticed that learners had specific problems while interacting with simulations. De Jong and van Joolingen (1998) categorized these observations and associated the revealed problems with processes in discovery learning. The major difficulties learners had included failure to generate new hypotheses or adapt hypotheses based on gathered information; a tendency to search for information that confirms, rather than disconfirms, a hypothesis; designing experiments that do not allow one to draw valid conclusions; insufficient experimentation behavior; difficulty interpreting experiment data; and unsystematic planning and monitoring during the learning process. The researchers argued that these intrinsic problems prevented simulation-based learning from reaching its full potential.

Because of the problems learners experienced in simulation-based environment, many researchers and educators contended that learners should be guided in their exploration, and some further suggested that learning should be structured so that learners are led through different stages of the discovery process rather than being left alone. As a result, more recent empirical work on simulation-based learning focused more on the effects that different types of instructional support had on interactive learning. The cumulative results from research along this line showed that providing assignments
(including questions or exercise) and structuring the simulation environment by dividing up the learning process were among the most effective measures that had positive effects on learning outcomes (de Jong & van Joolingen, 1998).

A study by Linn and Songer (1991) demonstrated the advantage of a structured environment over an unstructured environment in interactive learning. The study was a series of four classroom experiments in which middle school students learned topics in thermodynamics in an interactive learning environment. The same procedure was followed in all four experiments. First the teacher presented a question concerning a specific aspect of the subject matter. Then pairs of students conducted a controlled experiment as guided by the teacher. After the experiment was completed, students prepared reports and presented the results in a group discussion. The four experiments differed in how the interactive learning process was structured.

In the first two experiments, interactive learning was unstructured in that no additional activities were required of the students while they conducted the experiments in the interactive environment. The results from the two experiments showed that, although students gained some understanding of the impact of different variables on heating and cooling, they made little progress on developing integrated understanding of heat energy and temperature.

In the third experiment, two types of structure were posed on interactive learning. In the first, observation condition, students observed what happened in their experiments and recorded experiment logs. In the second, prediction condition, students predicted experiment outcomes based on previous results and, once the experiment was complete, reconciled their predictions with the actual outcomes. The results of this experiment
showed that both observation and prediction increased students' ability to integrate understanding of heat energy and temperature. However, observation was helpful primarily for students who had some understanding of the underlying principle prior to learning, while making predictions benefited students at all levels of initial understanding.

In the fourth experiment, students' activities during interactive learning included both observation and prediction. Compared to the unstructured interactive learning in the first two experiments, from two to four times as many students in this experiment were able to reach performance criteria on a test. These researchers concluded that in interactive learning, explicit measures should be taken to motivate students to construct understanding.

The different learning outcomes in structured and unstructured interactive learning were explained by some researchers from a cognitive load perspective. Sweller and colleagues (Sweller, van Merrienboer, & Paas, 1998) distinguished between two types of cognitive load associated with a learning task. The first type, the intrinsic cognitive load, is the inherent aspect of the mental task that is essential to achieve understanding and learning. The second type, the extraneous cognitive load, is associated with the way the instructional material is learned and may add to the intrinsic cognitive load. Due to the learner's limited capacity of working memory, inappropriate presentation of learning material or inappropriate learner activities can cause cognitive overload, leading to reduced capacity to solve problems. Therefore, learning is most effective and efficient if the unnecessary cognitive load is kept to a minimum.
Tuovinen and Sweller (1999) examined the cognitive load associated with discovery learning and worked examples in the context of learning to use a database program. In the experiment, the discovery learning group explored the database software by performing a number of tasks following printed instructions. They received the flowchart of steps involved in the tasks as an aid. The worked-examples group was directed to work through the problems in the printed practice instructions. They first read the initial problem and its worked-out solution, then interacted with the software with the same aid of flowchart. Participants in both learning conditions recorded the mental effort required to complete the tasks using a Likert scale. The results of the experiment showed that participants with no previous domain knowledge benefited substantially from worked examples in comparison to exploration. Their self-reported mental effort was significantly lower than that in the exploration group, and their performance on a related test was significantly better. However, for participants who were familiar with the database domain, the learning conditions did not make significant difference. Tuovinen and Sweller argued that this was because the exploration group was able to draw on their existing domain schemas to guide their learning. They further reasoned that the difference on mental effort ratings suggested that the advantage of worked examples was due to different cognitive load in the two learning conditions.

Computer-based simulations have been increasingly incorporated into statistical curricula at various educational levels. Although there have been only a limited number of empirical studies researching the effect of interactive learning in statistics education (Mills, 2002), the existing body of research did provide helpful insight into the potentials and challenges of simulation-based learning in this particular subject domain.
Statistical reasoning is the culturally prescribed way to think about events of a highly probabilistic nature in everyday life. Yet it is a well-documented phenomenon that, when making inductive reasoning, people often deviate from statistical principles in favor of non-statistical heuristics (Kahnman, Slavic, & Tversky, 1982). In the past, various training regimes were devised to improve people's performance on statistical reasoning tasks. These regimes, grounded on different assumptions of human reasoning, followed different instructional approaches such as the pragmatic-implications approach (Hilton, 1995), the heuristics and biases approach (Beyth-Marom & Dekel, 1983), the abstract rules approach (Nisbett, Fong, Lehman, & Chang, 1987), and the adaptive algorithm approach (Cosmides & Tooby, 1996). However, in almost all the approaches, teaching was very much of an expository nature, and training seldom yielded significantly positive effect on learners' actual performance.

Sedlmeier (1999) observed these limitations and designed an experiment in which learners interacted with a computer simulation to learn the effect of sample size on sampling outcomes. Participants' solutions of two well-known sampling distribution problems, the maternity-ward question and the word-length question*, were compared to responses from no-treatment controls. The results showed that simulation-based learning improved participants' statistical reasoning. The effect size attributed to the interactive learning across the two tasks was $r = .24$. More importantly, on the word-length problem, the correctness of solution was substantially correlated with the number of simulation runs learners had conducted during learning. The more a participant interacted with the program, the better his/her performance was. Although this study did not specifically

* A version of the maternity-ward problem is provided in Appendix C, Item 5. The original word-length problem was provided in Kahneman and Tversky (1972).
compare interactive learning with expository learning, such outcome, when compared to
the effects of the various training regimes mentioned above, strongly suggested that
interacting with the computer simulation improved participants' reasoning.

Research on the effectiveness of simulation in statistical education has also
looked into the impact of structure in simulation-based learning. delMas and colleagues
(delMas, Garfield, & Chance, 1999) incorporated a simulation program on the concept of
sampling distribution into classroom use and assessed the effect of learning structure in
two studies. In their first study, students received ancillary instructional materials
containing clear directions to guide their exploration in the simulation. They were tested
on a set of diagnostic graphics-based measurement items before and after interacting with
the simulation, and their performances on the two tests were compared. The results
showed that, despite of an overall significant change from pretest to posttest, a substantial
number of students still had serious misunderstanding of the subject matter. In the second
study, the learning structure was improved so that students used the simulation to verify
their predictions about the simulation outcome measured on the pretest. The results
showed that students' posttest scores were significantly higher than that in the first study.
These researchers concluded that student activities in a simulation environment do not
necessarily lead to a sound conceptual understanding. However, interactive learning can
be effective if the learning is structured in a way to encourage conceptual change.

The cumulative findings from empirical research on simulation-based interactive
learning, in both statistics and other domains of science education, indicated that
interactivity can enhance learning but the structure of the learning process is critical to
the learning outcome. Proper learning support should be provided in a simulation environment in order to ensure good learning results.

In past research, simulation-based interactive learning and expository learning often used different types of content delivery medium. Understandably, expository learning usually involved more traditional medium types, such as textbooks, tutorials, or lectures. Research has shown that a computer simulation has advantages over these traditional medium types. For example, Lane and Tang (2000) found that a simulation-based demonstration of a statistical principle, the law of large numbers, was more effective than a textbook-based learning method. Therefore, in order to accurately evaluate the effect of learner control in a simulation-based environment, it is important to separate interactivity from the medium effect.

The current research was designed to: (1) compare simulation-based interactive learning with expository learning in the context of learning a complex statistical principle, the law of large numbers; and (2) investigate the effects of different structures on interactive learning. It consisted of three experiments. In the first experiment, simulation-based interactive learning was compared to a textbook-based learning method. In the second experiment, both the interactive and expository learning conditions used the same computer program to control for possible medium effect. In the third experiment, the results of interactive and expository learning in a simulation environment were compared at two different points in time to determine if there was any delayed effect. In different experiments, interactive learning possessed different structures. Overall, three types of learning support were examined. These included, in the order of decreasing structure, directive guidance, worksheets, and nondirective guidance.
2. EXPERIMENT 1

Experiment 1 was designed to compare the effectiveness of simulation-based interactive learning with that of textbook-based expository learning. Further, it aimed to investigate the effects of different learning structures on simulation-based interactive learning. The experiment included three treatment conditions. In the first two conditions, participants learned the concept of sampling distribution through interacting with a computer simulation, with the learning process structured in two different ways. In the first condition, participants received detailed step-by-step instructions on how to interact with the simulation program. This is called the directive guidance condition. In comparison, in the second condition, participants received a set of high-level questions to lead them through the learning process. This is called the nondirective guidance condition. In the third condition, participants learned the material through reading a textbook chapter. Finally, a control group was also included to provide baseline information on people's intuitive statistical reasoning of everyday-life situations.

2.1 Method

2.1.1 Participants

Eighty-one Rice University undergraduate students participated in this experiment in exchange for extra credits for their psychology classes. Their age ranged from 17 to 23 years old. Participants came from a variety of academic disciplines, but none of them had taken any statistics class before. The number of participants in the directive guidance, nondirective guidance, and textbook-based learning conditions were 23, 22, and 20, respectively. The control condition had 16 participants.
2.1.2 Materials

2.1.2.1 Learning Materials

For the textbook-based expository learning condition, a textbook chapter was composed of parts adapted from several well-written textbooks in statistics (Lowry, 1989; Rosenberg, 1990; Spatz, 1993). The chapter covered key concepts such as sample, sample size, frequency distribution, sampling distribution, and included a formal statement of the law of large numbers. As in a typical statistics textbook, the chapter was largely text but also included graphics, mathematical formulas, and a table detailing a hypothetical sampling distribution. The material was prepared in print format as well as recorded on an audio tape.

The computer simulation used for interactive learning was adapted from the Rice Virtual Lab in Statistics (Lane, 1999). Figure 1 shows the user interface of the program. For both interactive learning conditions, a script was written to introduce the same statistical concepts as covered in the textbook chapter, but the concepts were all explained in reference to a concrete example. The script did not reveal the law of large numbers until after the participants had finished interacting with the simulation. The script was tape-recorded and played back in the experiment to accompany the simulation.
Figure 1. User interface of the Sampling Distribution simulation. Of particular relevance to the current experiment was the effect of sample size on the dispersion tendency of the resulting sampling distribution.

Participants in the two interactive learning conditions were provided with additional material to guide them through the learning process. Those in the directive guidance condition received step-by-step instructions on how to set the parameters for the simulation. Participants in the nondirective guidance condition were given 10 high-level questions concerning various aspects of the statistical property of the simulation. Both the step-by-step instructions and the high-level questions were organized in a way that matched the content structure in the textbook chapter. They were presented in paper format. Excerpts of both types of guidance are provided in Appendix A.
2.1.2.2 Testing Materials

Two tests were used in this experiment. The first test consisted of 10 multiple-choice questions designed to tap into participants' knowledge of the subject matter. The questions were adapted from a high school Advance Placement test in statistics and the aforementioned statistical textbooks. The questions with their respective alternates were phrased in statistical terms. This set of questions is provided in Appendix B.

The second test was a transfer test including 10 scenarios that were of an everyday-life nature. The scenarios were adapted from past research on statistical reasoning (Fong, Krantz, & Nisbett, 1986; Jones, 1990). Participants were asked to reason about these scenarios. All 10 scenarios were embedded in contexts very different from those encountered during the learning stage. Among the 10 items, 7 scenarios pitted a larger sample against a smaller sample. These were the critical items that required applying the law of large numbers in reasoning. The remaining three scenarios served as filler items. In the experiment, participants in the three treatment conditions were informed that not all problems were relevant to what they had just learned, and that they should use their judgment carefully. As it did not apply to the control condition, this message was omitted from the instructions for the control group. The 10 scenarios in the transfer test are provided in Appendix C.

In addition to the two tests, a questionnaire was constructed to measure participants' subjective responses to their learning experience. The questionnaire consisted of 10 items. Among them, five items focused on participants' perception of learning effectiveness, and the remaining five items were centered on participants' affect
towards the specific learning method. Participants responded by rating each item on a 5-point Likert scale. This questionnaire is provided in Appendix D.

2.1.3 Procedure

The experiment was conducted in small groups of two to five participants. The groups were randomly assigned to one of the four experimental conditions.

Procedure for the treatment conditions differed from that for the control condition. In the three treatment conditions, a 12-minute paper-and-pencil Wonderlic Personnel Test* was first administered when the experiment started. After finishing this test, participants read two scenarios that pitted a smaller sample against a larger sample. Each scenario came with three specific questions about the sampling outcome as a result of different sample sizes. Participants were asked to think about these questions but were not required to give answers. This took five minutes. Then they learned the material as prescribed by the treatment condition (described below). Testing began immediately after the learning stage. Participants first took the multiple-choice test for six minutes. Then they were given time to work on the transfer test. The time allotted to this test was not limited, but the majority finished in less than an hour. After the transfer test, participants filled out the questionnaire, and the experiment was concluded. Overall, the testing stage lasted no more than one hour.

In both simulation-based interactive learning conditions, participants learned the material in two steps. First, they listened to the script and observed the simulation

* Wonderlic Personnel Test is a 50-item omnibus test of general cognitive ability. The items provide a broad range of problem types, including analogies, analysis of geometric figures, disarranged sentences, definitions, and so on. The test generates one final score which is the sum of the correct answers. In the national norm, the first, second, and third quartiles of test scores are 16, 21, and 26, respectively.
program for eight minutes to learn the basic concepts. Occasionally they were asked to set parameters for the program so as to make the simulation progress in accordance with the script. During the second step, the directive and nondirective guidance groups followed different instructions and interacted with the simulation on their own for 20 minutes. The interaction sequence was logged in a file for each individual. After that, participants in both conditions listened to the concluding part of the script summarizing the law of large numbers.

In the textbook-based expository learning condition, participants read the chapter in print while listening to the accompanying audio tape. The tape paused at a number of points to allow participants to digest graphs, formulas, and examples in the chapter. It took participants 22 minutes to finish reading the material.

Procedure in the control condition was different from that in the treatment conditions. In this condition, participants first took the Wonderlic Personnel Test, then immediately proceeded to answer the transfer questions without going through the learning stage.

2.1.4 Data Analysis

Participants’ log data were analyzed to reveal the number of simulation runs during the interactive learning process. A run was defined as changing of a critical parameter in the simulation program, such as the population or sample size. This measure was obtained for both the directive and nondirective guidance conditions, and the two results were compared against each other.
Participants' reasoning in the transfer test was transcribed and scrambled so that all participants' responses to individual questions were gathered and coded together. Doing so concealed the experimental condition and also reduced the bias that could arise if the answers from each individual were scored all at once. For the seven critical items, each answer was rated on a 3-point scale based on how sophisticated the reasoning was in a statistical sense. This coding scheme is described in the next section. For the three filler items, an answer was rated as correct as long as it did not make reference to the law of large numbers. However, if this did happen, the answer was coded as a false alarm. Two coders independently rated the responses from a random sample of 10 participants (12% of all), and agreed on 87% of the answers. This inter-rater consistency was considered acceptable. All the discrepancies on ratings were no more than one point apart, and were subsequently reconciled. Then one of the two coders continued to code the remaining responses.

A participant's overall statistical reasoning score was obtained by averaging across this individual's ratings on all seven transfer questions. In addition, an individual's performance on the multiple-choice test was also obtained by calculating the accuracy across all 10 questions. Both measures were then analyzed using the Analysis of Covariance (ANCOVA) procedure. Participants' Wonderlic Personnel Test scores were taken as an index of their general cognitive ability, and were used as the covariate in the ANCOVA procedures.

The items in the questionnaire fell into two categories: perceived effectiveness and affect. Therefore, participants' responses were averaged within each category and analyzed with the Analysis of Variance (ANOVA) procedure.
2.1.5 Transfer Question Coding Scheme

Each response to a transfer question was classified into one of the following three categories on a 3-point scale:

1. An entirely deterministic response. In this category, there was no reference to statistical concepts, such as randomness, variance, or sample size, at all. Regarding the slot machine question (Appendix C, item 1), a sample answer in this category would be: The chances of winning on either machine are the same regardless of the outcome. Certain isolated events do not make a positive conclusion.

2. A poor statistical response. Responses in this category made use of certain statistical concepts, but were either incomplete or incorrect. A sample answer would be: I don’t think a conclusion can be made on just one trial. The old man has been playing longer, so he has more experience with the machines anyway. It just depends on how many times the game is played and the outcomes of the previous player.

3. A good statistical response. This would be an answer that correctly identified the sampling elements in the scenario as well as applied some form of the law of large numbers. A typical answer in this category would be: If the old man has tested the two machines for years and Keith has only played them for a few minutes, then the old man’s observations will probably reflect the average more closely. The more trials you run (years vs. minutes), the closer you will get to the real average.
2.2 Results

Participants in this experiment had relatively high cognitive ability. Their average Wonderlic test score was 29.1, with a standard deviation of 5.0. In comparison, the average score in the national norm is 21.1, with a standard deviation of 7.1.

Learner Interaction with the Simulation

For the two interactive learning conditions, the number of simulation runs represented the frequency of changing critical parameters in the simulation program during the learning process. Figure 2 shows the results when log data were analyzed for both conditions. As a group, participants who were guided with a set of high-level questions interacted with the simulation more than those who followed the step-by-step instructions. On average, participants in the nondirective guidance condition had 23 simulation runs (SD = 9.41) while those in the directive guidance condition had 16 runs (SD = 7.48). The difference was significant, \( t(42) = 2.75, p = .009 \).

![Box plot of simulation runs](image)

Figure 2. Frequency of interacting with the simulation as a function of learning structure in Experiment 1.
Knowledge Test

Figure 3 shows participants' performances on the multiple-choice test as a function of learning condition after adjusting for their Wonderlic test scores. The boxplots depict a varied degree of variability in accuracy in the three conditions. Performances in the two interactive learning conditions were more consistent than that in the textbook-based expository learning condition. The standard deviations of accuracy for the directive guidance, nondirective guidance, and textbook conditions were 14.4%, 15.8%, and 23.4%, respectively. Levene's test of homogeneity of variances showed that the variances in the three conditions were significantly different, $p = .010$.

![Boxplot](image)

Figure 3. Accuracy on the statistical knowledge test in the three learning conditions in Experiment 1.

The accuracy data were analyzed through ANCOVA, with participants' cognitive ability (i.e., Wonderlic test score) as the covariate. There was a significant effect of cognitive ability on the test performance, $F(1, 59) = 19.65, p < .0001$. Participants who scored higher on the Wonderlic test were more accurate answering the questions than did
those with lower Wonderlic test scores. The interaction between cognitive ability and learning condition was not significant, $F(2, 59) < 1$. Participants who learned the material through interacting with the simulation had higher response accuracy than those in the textbook-based expository learning condition. Both the directive and nondirective guidance groups had an average accuracy of 71.7%, while the textbook-based learning condition had an average accuracy of 66.2%. The differences, however, were not significant, $F(2, 59) < 1$.

Transfer Test

Participants' statistical reasoning on the transfer test was first analyzed as an Experimental Condition (4) x Scenario (7) ANCOVA with repeated measures on the last variable. There was no evidence that the scenarios were of significantly different difficulty, $F(6, 480) < 1$. The interaction between scenario and experimental condition was also not significant, $F(18, 480) < 1$.

Figure 4 displays the overall statistical reasoning scores on the transfer test for the four experimental conditions. On the 3-point scale, the average reasoning scores for the directive guidance, nondirective guidance, and textbook-based learning conditions were 2.06 (SD = 0.48), 2.18 (SD = 0.38), and 2.24 (SD = 0.41), respectively. Participants in the control condition had an average of 1.83 (SD = 0.37). The effect of cognitive ability on statistical reasoning was significant, $F(1, 73) = 5.59, p = .021^*$. Participants who had higher Wonderlic test scores also had higher reasoning scores on the transfer test. There

---

* Cognitive ability was found to have a significant effect on participants' performances on both the multiple-choice test and the transfer test in all three experiments in the current research. For the reason of brevity, ANCOVA results on the effect of cognitive ability are omitted from the rest of this report.
was no evidence of a significant interaction between cognitive ability and experimental condition, $F(3, 73) < 1$. The differences on performance variability among the four conditions were not significant, $p = .376$. Dunnett's test showed that the average reasoning scores in the nondirective guidance and textbook-based learning conditions were significantly higher than that in the control condition (both $ps < .05$). Pair-wise comparisons among the three learning conditions revealed no significant differences (all $ps > .05$).

![Figure 4](image-url)  

**Figure 4.** Statistical reasoning as a function of experimental condition in Experiment 1.

Participants reasoning on the three filler items had 19 (7.8%) false alarms out of a total of 243 responses. The false alarms were roughly evenly distributed in the three learning conditions, with the textbook-based, directive and nondirective guidance groups each made 5, 7, and 7 false alarms, respectively.
Responses to the Questionnaire

Participants' responses to the questionnaire are summarized in Table 1. Overall, participants in this experiment were positive about the effectiveness of all three learning methods. On a 5-point scale, the average ratings in the directive guidance, nondirective guidance, and textbook-based learning conditions were 3.9, 3.9, and 3.7, respectively. The differences were not significant, $F(2, 61) < 1$. By contrast, the groups differed to a certain extent on the affective measure. Participants in the interactive learning conditions had neutral to slightly positive affective responses. Their average ratings on this measure were 3.1 in the directive guidance condition and 3.0 in the nondirective guidance condition. The textbook-based learning group, on the other hand, leaned towards the negative side of the scale, with an average rating of 2.6. The differences were significant, $F(2, 61) = 3.23, p = .046$. Pair-wise comparisons further showed that the rating in the directive guidance condition was significantly higher than that in the textbook-based learning condition, $p = .041$.

<table>
<thead>
<tr>
<th>Subjective measure</th>
<th>Learning condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Directive guidance</td>
</tr>
<tr>
<td>Perceived effectiveness</td>
<td>3.87</td>
</tr>
<tr>
<td></td>
<td>(0.67)</td>
</tr>
<tr>
<td>Affect</td>
<td>3.12</td>
</tr>
<tr>
<td></td>
<td>(0.96)</td>
</tr>
</tbody>
</table>

Note. Values in parentheses denote standard deviations.
2.3 Discussion

One of the major goals of the current research is to examine the effect of learning structure on simulation-based interactive learning. In Experiment 1, interactive learning was structured in two different ways so that participants received either directive or nondirective guidance when they explored the Sampling Distribution simulation to discover the underlying statistical principle, the law of large numbers. The step-by-step interaction procedures in the directive guidance condition were well structured to ensure that the participants were exposed to different simulation scenarios so that they could potentially learn the various aspects of the subject matter. However, because participants were expected to follow the already specified interaction sequences, there was a risk that they might focus solely on executing the instructions instead of actively processing the information. By contrast, the high-level questions in the nondirective guidance condition created a less-well defined structure, thereby giving learners more control over the learning process. These questions could potentially promote information processing and facilitate learners conducting experiments in the simulation environment. On the other hand, these experiments were seldom complete or conducted in the optimal sequential order. As a result, learners in this condition might not be able to learn the full scope of the content given the time constraint in the current setting.

The results of Experiment 1 show that participants in the two interactive learning conditions differed in the degree to which they interacted with the simulation program. Participants who received directive guidance followed closely the step-by-step instruction. During the 20-minute period when they interacted with the simulation, these participants conducted an average of 16 simulation runs, which closely matched the
prescription (13 runs) in the instructions. In comparison, participants led by the nondirective guidance interacted considerably more with the program, resulting in an average of 23 simulation runs. Compared to the step-by-step instructions, the challenge of making sense of these high-level questions and finding solutions by designing and implementing various experiments in the simulation environment was much more difficult in the nondirective guidance condition. Therefore, this result indicates that in simulation-based interactive learning, nondirective guidance promotes learner to explore the simulation environment more than directive guidance does.

In Experiment 1, participants in the two interactive learning conditions had the same level of performance on the knowledge test. Both groups' response accuracies on the multiple-choice test, 72%, were moderate considering the complexity of the subject matter and the duration of learning. Although all the specifics about the underlying statistical principle were covered in the step-by-step instructions, they were only implied in the high-level questions in the nondirective guidance condition and hence were dependent on each individual participant's interaction with the simulation. The equal performance in this learning condition, therefore, indicates that learners' self-directed interaction was as effective as the purportedly optimal instructions in the directive guidance condition.

The results from Experiment 1 also show that, to a limited extent, participants in both interactive learning conditions were able to apply the law of large numbers to novel situations. Their reasoning of real-world scenarios was moderately sound in a statistical sense. Although the two groups did not differ significantly on this transfer test, the nondirective guidance group slightly outperformed the directive guidance group. As
interactive learning in the nondirective guidance condition required the participants to engage in more complicated information processing, their better performance on this test suggests that enhanced learner control in a less structured interactive learning environment might have a beneficial effect on transfer of learning. Therefore, even though these participants' declarative knowledge about the subject matter remained the same as that of the directive guidance group, they were better at identifying the statistical nature of the transfer questions and so were more successful on this test.

Besides the effect of learning structure, another major goal of Experiment 1 was to compare simulation-based interactive learning with textbook-based expository learning. To this end, simulation-based interactive learning as a whole did not show definitive advantage over textbook-based expository learning on participants' test performances. In the current experiment, both interactive learning conditions had higher response accuracy on the knowledge test than did the textbook group, but were slightly outperformed by the latter group on the transfer test. The differences on both tests, however, were not significant. Despite this overall result, the interactive learning groups were significantly different from the expository learning condition on two measures. First, their performances on the knowledge test were more consistent while the performances in the textbook group had a significantly larger variability. Because both interactive learning conditions were also more accurate on this test, the increased consistency of test performances suggests that, regardless of the structure of learning, interacting with the simulation did improve participants' understanding of the subject matter at an individual level. The second significant difference between interactive and expository learning conditions was reflected in participants' subjective responses.
Although both were perceived to be equally effective, participants in the interactive learning conditions were more positive than the textbook group on the affective measure. This result is consistent with previous research finding that giving learners control over the learning process has a positive impact on learners' motivation.

The result from Experiment 1, that the interactive learning groups had performed at the same level as the textbook group overall, does not lend much support to the hypothesis that learner control promotes better learning than expository learning. A number of reasons might have contributed to this result, among which three seem to be the most probable. The first explanation had to do with the particular setup in each learning condition. While the textbook group learned the material by reading the print and listening to the tape, the two interactive learning conditions relied solely on the audio to learn the basic statistical concepts. Thus, it was more difficult for these participants to retain the information. The second possibility had to do with the extra mental effort associated with interactive learning. Participants in general were unfamiliar with the method of self-initiated interactive learning. In this specific setting, they were required to take on two tasks, one to discover the underlying statistical principle, the other to learn the specifics of controlling the simulation program. The latter task was especially challenging for those participants who received nondirective guidance. By contrast, learners in the textbook condition only need to process the already well-organized material. Although the duration of the interactive learning was six minutes longer than that of textbook-based learning, it did not seem to compensate for the greater mental effort associated with interactive learning. The third explanation as to why interactive learning did not generate better learning outcome is that learners' aptitudes might have
come into play in deciding the effect of different treatments. Specifically, as the
participants in this experiment had high cognitive ability, the advantage of interactive
learning might not manifest even if it could exist for learners with lower cognitive ability.
As aptitude-treatment interaction is a topic that merits a more thorough treatment, it will
be discussed in the next section.

2.3.1 Aptitude-Treatment Interaction in Interactive Learning

Aptitude-treatment interaction refers to the selective effectiveness of an
instructional method on learners who differ on measurable characteristics. Aptitudes are
defined as initial states of persons that influence later developments, given specified
learning conditions (Snow, 1992). Therefore in educational research, aptitudes include a
broad range of personal attributes, such as knowledge, skills, abilities, personality,
motivation, interests, attitudes, and so forth.

The notion of aptitude-treatment interaction is consistent with the emphasis on
individual differences in the cognitive learning paradigm. Some psychologists asserted
that aptitude-treatment interactions are ubiquitous in educational settings (Cronbach &
Snow, 1977). When a group of students learn with the same instructional method, a
variety of individual differences among these learners play crucial mediating roles in the
process, leading to differentiated outcomes. Therefore, no single teaching method is best
for all learners. Glaser (1972) further pointed out that the emphasis of research on
aptitude-treatment interaction is to identify aptitude measures that are useful for selecting
learning methods to maximize individual attainment of specific educational objectives.

Historically, aptitude-treatment interaction has been very difficult to find in
empirical research. This was often caused by methodological problems such as
insufficient sample size, poor research design, or unanticipated interactions across settings. Even in studies where the interactions were documented, the results were often inconsistent. For example, early research on the interaction between learners' ability and the structure of discovery learning favored the conclusion that more able learners do better with less structure, and less able learners do better with more (Snow, 1989). However, there was evidence that in some settings, low ability learners benefited more from discovery learning than high ability learners (e.g., Corman, 1957). Despite these inconsistent results, evidence about aptitude-treatment interaction in interactive learning has been accumulating.

Individual differences are an important consideration in interactive learning. In many of the empirical studies on this topic, interactive learning was situated in a hypermedia environment. The personal attributes examined in quantitative research covered a broad range but could be classified into two categories: ability and learning style. For example, Recker and Pirolli (1995) reported an experiment where they found that interactive learning had differentiated effect on learners with different ability. In their experiment, students learned Lisp programming in one of two learning environments: hypermedia, in which random information access was allowed, and a linearly organized electronic text. Participants' ability was defined as their level of knowledge, determined by their performance on an initial problem-solving exercise. The results of the experiment showed no significant difference between the two learning environments in general. However, when students' ability was taken into account, the data showed that high-ability learners benefited the most from the hypermedia environment while low-ability learners' performance decreased after the interactive learning. These researchers
reasoned that the low-ability students might have been overwhelmed by the amount of learner control in the hypermedia environment, thereby not being able to take advantage of this learning environment as high-ability learners did.

In addition to ability, learning style has also been demonstrated to play a role in determining the effectiveness of interactive learning. Learning style refers to an individual's distinct approach to learning. It posits a learner on a continuum between two extreme traits such as passive and active, analytic or holistic, and so on. In a literature review on hypermedia learning, Dillon and Gabbard (1998) noticed that learner characteristics were most frequently categorized along three continuums: field dependent/field independent, passive/active, and deep/shallow processors. For example, in a study by Beishuizen, Stoutjesdijk, and van Putten (1994), participants were measured with the Inventory of Learning Styles and categorized as either deep or surface processor. Deep and shallow processing refers to the degree of structural or surface analysis learners make when they encounter new information. The participants were randomly assigned to one of three hypermedia environments. Interactive learning was structured in different ways in the three environments. In the first two, participants received hints about study strategies or content structuring help. In the third condition, students received no such help. The researchers analyzed both the learning process and the outcome measures, and found no significant main effects for either learning structure or processing style. However, they identified interactions suggesting that content structuring help benefited surface processors but had a detrimental effect on deep processors.

Based on their literature review, Dillon and Gabbard (1998) made the following observations regarding aptitude-treatment interaction in interactive learning. First,
increased learner control differentially benefits learners with different ability. Specifically, lower ability students have the greatest difficulty with hypermedia learning. Second, the inconsistent results in literature concerning the effect of interactive learning may be explained with reference to the interaction of learning style.

There has been some evidence from empirical research indicating aptitude-treatment interaction in learning statistics in a simulation-based environment. Specifically, it has been found that in such settings, a learner's cognitive ability plays an important role in determining the effect of interactive learning. Although the findings from different studies are not entirely consistent, the weight of evidence is pointing in the direction that learners with low cognitive ability benefit more from interactive learning than those with high cognitive ability.

Weir, McManus, and Kiely (1991) reported a study in which interactive learning was compared to lectures in learning statistical concepts. In the context of normal teaching, two matched groups of undergraduate students learned the concepts of standard error and F-distribution. Each group learned one of the two concepts with a simulation and the other concept through a regular lecture. The students were tested on their knowledge of both concepts. The results showed that the F-distribution simulation had a differentiated effect on learning. Students with high cognitive ability performed at the same level regardless of the learning method. By contrast, students with low cognitive ability had significantly higher test performance in the interactive learning condition than in the lecture condition. In a follow-up study with a new group of students, the standard error simulation was modified and used together with the F-distribution simulation. The results showed that interactive experience with both simulations resulted in better
learning for the students than the lectures. Although high-ability learners in interactive learning also improved their test score, the benefit of simulation was more substantial for those with low cognitive ability.

In an interactive learning environment, the structure of learning is instrumental to the effectiveness of the program. Veenman and Elshout (1995) conducted a research in which they manipulated the structure of a statistical simulation and examined the impact on learners with high or low cognitive ability. In the study, participants learned the formula for correlation by experimenting with datasets. The learning environment was either structured in that the datasets were presented in two by two diagrams, or unstructured such that data were presented as a row of numbers that had to be reorganized by the learners themselves. Participants' meta-cognitive ability was also measured by analyzing their verbal protocols during interactive learning. Learning was assessed on both declarative and procedural domain knowledge. The results showed that the structure of the learning environment did not affect learners with high cognitive ability regardless of their meta-cognitive skillfulness. However, low cognitive ability learners with poor meta-cognition had enhanced performance in the structured environment when compared to the unstructured learning environment.

On other occasions, research results did not generate a clear-cut answer regarding the effectiveness of interactive learning but rather provided some suggestions that aptitude-treatment interaction might have played a role in determining the learning outcome. For example, Chen (1999) had college students learn the law of large numbers in three different ways. The first, interactive-simulation group learned through playing a card-dealing simulator. The second, passive-simulation group learned by watching the
interactive sequence recorded in the first group. The third, expository group read a story set in the same card-dealing context. The three groups' performances on a transfer test were compared. Although interactive learning did not lead to better performance overall, participants' reasoning of one particular type of transfer question showed that interactive simulation was the most effective method for lower cognitive ability learner while text was the most effective method for learners with high cognitive ability. The interaction between learning method and cognitive ability approached significance ($p = .074$).

In view of the evidence from the literature, that interactive learning could benefit low ability students in learning statistics more than high ability learners, the results of Experiment 1 might have reflected its limitations in research methodology. First, the participants in this experiment had high cognitive ability, with an average Wonderlic score the same as that of a college graduate in the national norm. It was possible that interactive learning in this particular setting might selectively affect learners on the lower, rather than the higher, end of the ability scale. Second, the ways that interactive learning was structured in the current experiment, the directive and nondirective guidance, might not provide the optimal learning support in the current context. Therefore, Experiments 2 and 3 were designed to address these two considerations.

3. EXPERIMENT 2

Experiment 2 was designed to test if learner control could lead to better learning outcome than expository learning when proper learning structure was provided in interactive learning. It also aimed to examine if there was any aptitude-treatment
interaction in the current learning context, such that interactive learning could be more effective for participants with lower, rather than higher, cognitive ability.

The experiment employed a 2 x 2 (Learner Control x Prediction) factorial design. On the first factor, participants learned the material either through interacting with the simulation program or by watching a simulation demonstration. On the second factor, participants either predicted the simulation outcomes at certain points during the learning process or proceeded without making predictions. A control group was also included to provide baseline information on people’s reasoning of everyday-life situations.

There were a number of methodological improvements in the current experiment compared to Experiment 1. First, while different content delivery media were used in the interactive and expository learning conditions in Experiment 1, the same simulation program was used in both conditions in the current experiment. Second, the structure of interactive learning was improved so that now the learning sequence for those who interacted with the simulation was divided into three steps, and a worksheet was provided in each step as an aid for interacting with the program. The third improvement addressed the restricted range of cognitive ability in the previous experiment. In Experiment 2, participants were recruited from two universities that differed in their respective students’ academic achievement. The much bigger variation in participants’ cognitive ability allowed for increased statistical power in detecting any potential aptitude-treatment interaction in the current setting.
3.1 Method

3.1.1 Participants

One hundred and thirty-two undergraduate students were recruited from Rice University and University of Houston Downtown Campus to participate in the experiment. They received either extra credits for their psychology classes or a nominal amount of monetary reimbursement for their participation. Participants' age ranged from 16 to 48 years old, with a median of 19. They came from a variety of academic disciplines, but none of them had taken any statistics class before. The number of participants in the four treatment conditions ranged from 26 to 30. The control condition had 22 participants.

3.1.2 Materials

Learning Material

The Sampling Distribution simulation program was the same as in Experiment 1. The simulation was used in both the interactive and expository learning conditions in the current experiment.

For the expository learning condition, a script was written to accompany the simulation demonstration. The script first introduced key statistical concepts, then discussed three aspects of the sampling distribution in sequence: its shape, its statistical properties, and the effect of sample size on the statistical properties. Finally, the script gave a formal statement of the law of large numbers. The script was pre-recorded and played back in the experiment. The simulation software was programmed to
automatically display, in a time-locked fashion, the series of events prescribed in the script.

For the interactive learning condition, only two portions of the script, the introduction to basic statistical concepts at the beginning and the formal statement of the law of large numbers at the end, were used. The simulation program was kept intact. In addition, three worksheets were designed to advice participants on how to interact with the program. Each worksheet focused on one of the three aspects of the sampling distribution as mentioned above. The worksheets organized the simulation parameters in tabular formats. While the parameters for some simulation runs were predefined, participants were also asked to set the parameters on their own for other simulation runs. Once the parameters were set, participants were to interact with the program and find out the simulation result in each setting. A sample worksheet is provided in Appendix E.

**Testing Material**

Testing materials included multiple-choice test and transfer test. The multiple-choice test was the same as in Experiment 1. The transfer test consisted of six of the seven critical items and all three filler items from Experiment 1. In addition, three new transfer questions were added. Altogether, there were 12 scenarios (nine critical items and three fillers) in the transfer test. The newly added transfer scenarios are included in Appendix C. Finally, a questionnaire of eight survey items (all from Experiment 1) was used to measure participants' subjective responses to learning. Among the survey items, five were centered on perceived learning effectiveness and the remaining three tapped into participants' affective response.
3.1.3 Procedure

Small groups of two to six people were randomly assigned to one of the five experimental conditions. The overall procedure was similar to that in Experiment 1. When the experiment first started, all participants took the Wonderlic Personnel Test. Those in the treatment conditions then read the two sampling scenarios (same as in Experiment 1), and learned the material as prescribed by the treatment condition. Testing took place immediately after the learning stage. Participants first were given six minutes to take the multiple-choice test. Then they took the transfer test, with most of them completing the test in less than one hour. Finally, participants filled out the questionnaire. As in Experiment 1, the controls did not go through the learning stage and were tested only on the transfer questions.

The learning procedure differed in the four treatment conditions. In the expository learning condition, participants watched the simulation demonstration while listening to the playback of the accompanying script. This process took 23 minutes. In the interactive learning condition, participants first listened to the beginning portion of the same script for 10 minutes, then interacted with the simulation program on their own. They received the three worksheets one at a time. Each worksheet instructed participants to pay attention to one specific aspect of the simulation. Participants conducted simulation runs according to the specifications on the worksheet and observed the results. They were given 5, 5, and 10 minutes each to work on the three worksheets. Upon finishing a worksheet, the script briefly summarized the results of the simulation runs. Finally, after participants had gone through all three worksheets, the script gave a formal statement of the law of large numbers. These were the procedures for the non-prediction groups. For
the prediction groups, 10 questions regarding the simulation outcomes were interspersed in the demonstration script and the interaction sequence specified in the worksheets. Participants were given an extra 30 seconds to respond to each prediction question. The total learning time for these participants was the same as that for their non-prediction counterparts.

3.1.4 Data Analysis

Participants' answers in the transfer test were transcribed and scrambled to conceal the experimental conditions. All participants' answers to the same question were coded together. The answers were judged according to the coding scheme described in Experiment 1. Two coders each rated the answers from a random sample of 20 participants (15% of all), and agreed on 85% of the ratings. This inter-rater consistency was accepted. The discrepancies on ratings were reconciled, and one coder continued to rate the remaining answers.

Participants' overall statistical reasoning score and accuracy on the multiple-choice test were calculated in the same way as in Experiment 1. Both were then analyzed through ANCOVA, with participants' Wonderlic test score as the covariate.

Survey data were aggregated into the perceived effectiveness (five items) and affect (three items) categories, and were analyzed with ANOVA with a Learner Control (2) x Prediction (2) model specification.
3.2 Results

In the current experiment, participants' general cognitive ability had more variability than in Experiment 1. Their Wonderlic test scores ranged from 10 to 42, with an average of 23.6 (SD = 7.4). This average was much lower than that in Experiment 1 (29.1), and was close to the national average (21.1).

Knowledge Test

Participants' response accuracies on the knowledge test were first analyzed through ANCOVA, with treatment condition as the sole independent variable. There was no evidence of a violation of the assumption of homogeneity of regression slopes, $F(3, 102) = 1.03, p = .382$. The data were then analyzed with a Learner Control (2) x Prediction (2) ANCOVA procedure. The results of this test are provided in Appendix F.

Figure 5 shows participants' response accuracy as a function of learning condition. On average, the interactive learning groups answered 70.9% (SD = 17.7%) of the questions correctly, while those who watched the simulation demonstration had an average accuracy of 61.2% (SD = 20.7%). The difference was significant, $F(1, 103) = 6.71, p = .011$. Participants who predicted the simulation outcomes during learning had an overall accuracy of 68.7% (SD = 19.5%), higher than the accuracy of those who did not make predictions, 63.1% (SD = 20.0%). However, the difference between these two conditions did not reach significance, $F(1, 103) = 2.16, p = .145$. In addition, although the effect of prediction appeared to be stronger in the interactive learning condition than in the expository learning condition, there was no significant interaction between the two independent variables, $F(1, 103) = 1.33, p = .252$. 
Figure 5. Accuracy on the statistical knowledge test in the four treatment conditions in Experiment 2.

Figure 6 depicts response accuracy as a function of cognitive ability and learner control. For the convenience of illustration, participants were categorized into four cognitive ability groups based on their Wonderlic test score\(^*\). Figure 6 gives a strong hint that the advantage of interactive learning over expository learning was more prominent for participants with lower cognitive ability than for those with higher cognitive ability. However, ANCOVA results showed that the interaction between learner control and cognitive ability did not reach significance, \(F(1, 103) = 2.71, p = .103\). It is worth mentioning that in the ANCOVA procedure, cognitive ability was treated as a continuous variable (i.e., participants' original Wonderlic test score) rather than a categorical variable.

\(^*\)In the Low, Medium low, Medium high, and High cognitive ability groups, the ranges of participants Wonderlic test scores were lower than or equal to 16, 17 to 22, 23 to 28, and equal to or higher than 29, respectively. There were about equal numbers of participants (24, 23, and 24, respectively) in the first three groups, while the last group had considerably more (35).
Figure 6. Accuracy on the knowledge test as a function of cognitive ability and learner control in Experiment 2.

Transfer Test

Participants' statistical reasoning scores on the transfer test were first analyzed as an Experimental Condition (5) x Scenario (9) ANCOVA with repeated measures on the last variable. There was no evidence that the scenarios were of different difficulty, $F(8, 976) = 1.14, p = .337$, although there was a marginally significant interaction between scenario and condition, $F(32, 976) = 1.39, p = .074$. There was also a marginally significant interaction between experimental condition and cognitive ability, $F(4, 122) = 2.13, p = .081$.

Figure 7 shows participants' mean reasoning scores in the five experimental conditions. On the 3-point scale, the adjusted group average was 2.01 (SD = 0.44) for the demonstration only condition, 1.99 (SD = 0.46) for the demonstration and prediction condition, 2.09 (SD = 0.32) for the interaction condition, and 2.07 (SD = 0.49) for the
interaction and prediction condition, respectively. The control group had an average reasoning score of 1.69 (SD = 0.37). Dunnett’s test showed that the differences between all learning conditions and the control were significant (all $p < .05$).

![Boxplot of reasoning scores](image)

Figure 7. Statistical reasoning as a function of experimental condition in Experiment 2.

Participants’ reasoning scores were further analyzed with a Learner Control (2) x Prediction (2) ANCOVA procedure. The results showed that there was no significant main effect of either learner control, $F(1, 103) < 1$, or prediction, $F(1, 103) < 1$. The interaction between the two variables was also not significant, $F(1, 103) < 1$. However, there was a significant interaction between learner control and cognitive ability, $F(1, 103) = 6.01$, $p = .016$. The complete results of the ANCOVA test are included in Appendix F.

The interaction between learner control and cognitive ability is illustrated in Figure 8, which shows statistical reasoning score as a function of learner control for participants with different cognitive ability. Participants were categorized into one of four cognitive ability levels based on the criteria mentioned above. As can be seen in Figure 8, the benefit of interactive over expository learning was more prominent for participants
with lower cognitive ability than those with higher cognitive ability. Specifically, interacting with the simulation had significantly improved participants' statistical reasoning for those with medium low cognitive ability, $F(1, 18) = 5.14, p = .036$, but not for other cognitive ability groups (all $ps > .05$).

![Graph showing statistical reasoning scores for different cognitive abilities and learner controls.](image)

Figure 8. Statistical reasoning as a function of learner control and cognitive ability in Experiment 2.

Participants' reasoning on the three filler items was also analyzed. There was a total of 389 responses to the filler items. Participants in the treatment groups made 20 (5.1%) false alarms while the control group made none. The false alarms were almost evenly distributed in the four treatment conditions. In the expository learning condition, the no-prediction and prediction groups made five and six errors, respectively. In the interactive learning condition, the no-prediction and prediction groups made six and three false alarms, respectively. Fifteen of the 20 false alarms were made by participants in the
upper two cognitive ability groups, while those in the lower two groups had five false alarms.

**Responses to the Questionnaire**

Participants' responses to the questionnaire are summarized in Table 2. Regardless of the learning condition, the average ratings were mostly in the positive direction on the 5-point scale, indicating that participants generally had favorable opinions about their learning experience in this experiment.

The results also showed that participants had more positive reaction to interactive learning than expository learning. On the measure of self-perceived learning effectiveness, participants who interacted with the simulation had an average rating of 3.9 (SD = 0.82), while those who watched the simulation demonstration had an average rating of 3.5 (SD = 0.87). A 2 x 2 (Learner Control x Prediction) ANOVA test showed that the difference was significant, $F(1, 106) = 5.97, p = .016$. Similarly, on the affective measure, the interactive learning condition had an average rating of 3.29 (SD = 0.55), higher than the average rating of 3.11 (SD = 0.71) in the expository learning condition. However, the difference on this measure was not significant, $F(1, 106) = 2.08, p = .152$.

In comparison, making predictions about the simulation outcomes in learning did not have much impact on the subjective measures. On perceived learning effectiveness, participants who made no predictions had essentially the same rating ($M = 3.63, SD = 0.88$) as those who made predictions ($M = 3.67, SD = 0.86$). On the affective measure, making predictions resulted in a slightly more positive rating ($M = 3.29, SD = 0.61$) than
the non-prediction condition (M = 3.11, SD = 0.66). The difference, however, was not significant, $F(1, 106) = 1.91, p = .170$.

Table 2. Average ratings on subjective measures in the four learning conditions in Experiment 2

<table>
<thead>
<tr>
<th>Subjective measure</th>
<th>Expository learning</th>
<th>Interactive learning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No prediction</td>
<td>Prediction</td>
</tr>
<tr>
<td>Perceived</td>
<td></td>
<td></td>
</tr>
<tr>
<td>effectiveness</td>
<td>3.43</td>
<td>3.49</td>
</tr>
<tr>
<td></td>
<td>(0.88)</td>
<td>(0.88)</td>
</tr>
<tr>
<td>Affect</td>
<td>2.98</td>
<td>3.26</td>
</tr>
<tr>
<td></td>
<td>(0.71)</td>
<td>(0.69)</td>
</tr>
</tbody>
</table>

Note. Values in parentheses denote standard deviations.

3.3 Discussion

Experiment 2 was designed to investigate the effect of learner control and learning structure in simulation-based learning. Interactive learning was structured by providing worksheets as guidance for learner control. Compared to the directive and nondirective guidance in Experiment 1, the worksheets focused less on lower-level interaction sequence but contained sufficient information on how to explore the simulation in a systematic way. Thus, the current structure of interactive learning ensured that learners could progress in much the same way as in the expository learning condition while still having the flexibility of exploring the simulation on their own. A second improvement made in the current experiment is that the same simulation program was used as content delivery medium in both interactive and expository learning. Therefore, any differences in learning outcome between the two learning conditions could be more accurately attributed to the effect of learner control.
The results of Experiment 2 showed that learner control led to better learning outcome. This advantage was most evident on the knowledge test. The interactive learning groups, regardless of whether or not they made predictions during the learning process, had significantly higher accuracy on the knowledge test than participants who learned by observing the simulation demonstration. By contrast, although their statistical reasoning of the transfer questions showed a slightly higher level of sophistication when compared to the latter group, the advantage was minimal. In a way, this result replicated the finding from Experiment 1, where the benefit of interactive learning over textbook-based learning was reflected by participants’ performance on the same knowledge test. Thus, the results from both experiments suggested that learner control, defined in the current setting as the ability to manipulate statistical parameters to produce different simulation outcome, could have helped the participants to better understand the many facets of the subject matter.

Although interactive learning did not have an overall beneficial effect on participants’ reasoning of everyday problems, there was evidence from the current experiment that such beneficial effect did exist, but only for learners with a certain level of cognitive ability. On the transfer test, there was a significant interaction between learner control and participants’ cognitive ability. For learners with medium low cognitive ability, interactive learning resulted in a significant improvement in test performance. By contrast, the advantage of learner control over expository learning was much smaller in the low cognitive ability group and eventually diminished when participants’ cognitive ability extended towards the upper end of the scale. Results on the knowledge test corroborated this finding. Although the interaction between cognitive
ability and learner control only approached statistical significance \((p = .103)\) on the knowledge test, the effect of learner control on participants’ performance differed among participants with different cognitive ability. Compared to the expository learning condition, interacting with the simulation had the biggest advantage for those with medium low cognitive ability, the same group that showed the most benefit on the transfer test. The benefit was much smaller for learners with medium high cognitive ability, and was minimal for those resided on both ends of the ability scale. Thus, the two converging findings provided evidence for aptitude-treatment interaction in the current learning context.

The interaction between learner characteristics and instructional method has been a long-researched topic in psychology and education. However, empirical studies on this topic often generated inconsistent results. Although cognitive ability is the most frequently used aptitude in research, evidences from difference sources suggest that the interaction between cognitive ability and treatment can take different directions. There has been some evidence that cognitive ability plays an important role in interactive learning. For example, research on learning with hypermedia has shown that, while high ability students can take advantage of enhanced learner control, lower ability individuals have the greatest difficulty in such an environment (Dillon & Gabbard, 1998). In the current experiment, learner control was found to benefit particularly those with lower-medium level of cognitive ability. It is noteworthy that several other studies in simulation-based interactive learning have generated similar results. In particular, two experiments on the effect of simulation in learning statistics both showed that interactive learning was the most beneficial for students with low ability (Veenman & Elshout, 1995;
Weir et al., 1991). Thus, the cumulative evidence in this research area suggests that simulation-based interactive learning may be particularly suitable for students with less ability in learning statistics. Such findings have significant practical implications in educational settings.

In the current experimental design, asking learners to predict the simulation outcomes did not generate consistent results on different performance measures. Making predictions did not affect participants' performance on the transfer test. However, this additional learning support did produce quite sizable effect on the knowledge test. Specifically, the current data indicate that making predictions can be beneficial when combined with interactive learning. On the knowledge test, making predictions resulted in an increase of 10 percentage points on response accuracy in the interactive learning condition. As a comparison, in the simulation-based expository learning condition, those who made predictions had essentially the same response accuracy as those who did not. Because the predications were all with respect to the specifics of the learning content, it is not surprising that the effect of making predictions would manifest on the knowledge test rather than the transfer test. This result is consistent with previous research finding that making predictions is an effective way to provide structure in interactive learning (e.g., de Jong & van Joolingen, 1998). Thus, the current experiment lends further support to the claim made by other researchers that additional learning support should be provided in order for students to gain from their interactive learning experience (e.g., delMas et al., 1999).

Participants' subjective responses revealed an overall positive attitude in all four learning conditions. Overall, learner control did not have a differentiated impact on
learners' affective measure. However, the data did suggest that participants liked interactive learning more than expository learning when no additional learning support (i.e., making predictions) was provided. On the other hand, participants perceived learner control to be more effective than expository learning. This result was in accordance with their actual performance on the knowledge test.

In summary, results from the current experiment provide evidence that learner control can benefit learning in a simulation-based environment. The advantage of interactive learning over expository learning in this experiment was likely the result of increased variability that learners experienced while interacting with the simulation. The effect of learner control depended on the proper structure of the interactive learning process. There was also evidence that learner control might selectively affect learners with different cognitive ability.

4. EXPERIMENT 3

Experiment 3 was designed to examine the effect of interactive learning over an extended period of time. It also aimed to compare interactive learning with expository learning while controlling for the learning content. The experiment employed a 2 (Learning Condition) x 2 (Test Timing) factorial design. On the first factor, simulation-based interactive learning was compared to expository learning. Interactive learning was structured in a way similar to that in Experiment 2. The expository learning group, on the other hand, was yoked with the interactive learning group by passively observing the interaction sequence of the latter group. On the second factor, within each learning
condition, half of the participants were tested immediately after the learning stage, and the other half were tested after a one-week delay.

4.1 Method

4.1.1 Participants

Seventy six undergraduate students from Rice University and University of Houston Downtown Campus participated in this experiment in exchange for either class extra credits or a nominal amount of monetary reimbursement. Their age ranged from 17 to 43, with a median of 20. None had taken any statistics class before. The number of participants in the four experimental conditions ranged from 16 to 22.

4.1.2 Materials

Learning Material

The computer simulation program was the same as in the previous two experiments. The learning materials were similar to those used in Experiment 2. The materials included a 15-minute tape-recorded script for introducing relevant statistical concepts and a packet of three worksheets to guide learner control in the simulation program (Appendix E).

Testing Material

The multiple-choice questions, transfer test scenarios and questionnaire were the same as in Experiment 2.
4.1.3 Procedure

Small groups of two to six participants were randomly assigned to one of the four experimental conditions. All participants first took the Wonderlic test and read the two sample scenarios. During the learning stage, the interactive learning groups listened to the script to learn basic statistical concepts, and then interacted with the simulation. They received the three worksheets one at a time. For each worksheet, participants were given four minutes to interact with the simulation, during which time the interaction sequence (keystrokes) was logged. After finishing a worksheet, participants listened to the script briefly to learn what the simulation result should be. After going through all three worksheets, they learned the formal statement of the law of large numbers and the explanation to the previous two sample scenarios.

In the expository learning condition, the participants first listened to the script to learn the basic concepts. Then each participant was yoked with one individual in the interactive learning condition, and observed the reconstructed (from the log file) interaction sequence from that participant. To match the pair's cognitive ability as close as possible, a participant was yoked only with another participant from the same university. Participants in the yoked condition also received the three worksheets, and were informed of what to look for during each step. They learned what the simulation result should be upon finishing watching an interaction segment. Upon finishing all three, participants learned the law of large numbers and the explanation to the two sample scenarios, as did those in the interactive learning condition.
In both the interactive and expository learning conditions, about half of the participants were tested immediately after the learning stage, and the other half took the test after a one-week delay.

Due to attrition over the one-week delay, the interactive learning condition and the yoked, expository learning condition had unequal numbers of participants. Of the 76 participants who completed the experiment, 64 were from yoked pairs.

4.1.4 Data Analysis

Participants' reasoning of the real-world scenarios was transcribed and scrambled and then rated based on the coding scheme described in Experiment 1. Two coders separately rated the answers from a random sample of 10 participants (13% of all), and agreed on 88% of the answers. This inter-rater consistency was accepted. The discrepant ratings were reconciled, and one coder continued to rate the remaining responses.

Statistical reasoning scores and accuracies on the knowledge test of the 32 yoked pairs were first analyzed. On both measures, partial correlation between the two learning conditions, when controlled for Wonderlic test scores, showed no significant result. Therefore, the yoked pairs were treated as a between-subject design. Performance data were then analyzed through ANCOVA procedure, first for the 32 yoked pairs and then for all the 76 participants. Both methods generated similar results. The results reported below are based on data from all 76 participants.

Participants' subjective responses were aggregated into the categories of perceived effectiveness and affect, and were analyzed with ANOVA with a Learning Condition (2) x Test Timing (2) model specification.
4.2 Results

Participants' Wonderlic test scores ranged from 12 to 40, with an average of 25.54 (SD = 6.82). In the yoked pairs, the differences on Wonderlic test score between two participants ranged from -10 to 11. The mean difference across all pairs was 1.2 (SD = 5.30), with participants in the interactive learning condition fared slightly better than those in the expository learning condition.

Knowledge Test

Participants' performances on the multiple-choice test are shown in Figure 9. Three of the four experimental conditions had similar levels of response accuracy on this test. Average response accuracy in the interactive learning condition was 62.9% (SD = 17.6%) on the immediate test, and 62.5% (SD = 17.1%) on the delayed test. In the yoked expository learning condition, the average accuracy was 64.5% (SD = 17.3%) on the immediate test, and 70.9% (SD = 20.1%) on the delayed test. ANCOVA results showed that there was no significant effect of learner control, $F(1, 69) = 1.51, p = .223$; or test timing, $F(1, 69) < 1$. The interaction between the two variables was also not significant, $F(1, 69) < 1$. In addition, there was no evidence of interaction between cognitive ability and learner control, $F(1, 69) < 1$. 
Figure 9. Accuracy on the statistical knowledge test as a function of experimental condition in Experiment 3.

Transfer Test

Figure 10 shows participants’ performances on the transfer test in the four experimental conditions. The two learning conditions had similar reasoning scores on the immediate test. On the 3-point scale, the average reasoning score was 1.93 (SD = 0.46) for the interactive learning condition, and 1.94 (SD = 0.38) for the expository learning condition. Participants had slightly better scores on the delayed test, averaging at 2.00 (SD = 0.30) and 2.14 (SD = 0.47) for the interactive and expository learning conditions, respectively. The reasoning scores were analyzed through ANCOVA. There was no significant effect of learning condition, $F(1, 69) < 1$; or test timing, $F(1, 69) = 1.91$, $p = .171$. The interaction between the two variables was also not significant, $F(1, 69) < 1$. Furthermore, there was no evidence of interaction between cognitive ability and learner control, $F(1, 69) < 1$. 
Participants made 13 (5.7%) false alarms on the filler items, out of a total of 228 responses. The interactive and expository learning groups each made two false alarms on the immediate test, and four and five false alarms respectively on the delayed test.

**Subjective Responses to Questionnaire**

Table 3 summarizes participants' subjective responses to the questionnaire. Overall, participants in this experiment gave positive judgments regarding learning effectiveness, and had moderate affect toward their learning experience. The interactive and expository learning conditions had very similar responses on both subjective measures. On the 5-point scale, the average rating on perceived effectiveness was 3.62 (SD = 0.84) and 3.61 (SD = 0.86), and that on the affective measure was 3.05 (SD = 0.74) and 3.07 (SD = 0.72) for interactive and expository learning, respectively.
Subjective responses showed some fluctuation over the one-week delay. The average rating on perceived effectiveness was 3.68 (SD = 0.84) immediately after learning and 3.55 (SD = 0.86) after the one-week delay. The affective measure at the two points was 3.02 (SD = 0.70) and 3.11 (SD = 0.76), respectively. The fluctuation happened mostly in the expository learning condition. Responses in the interactive learning condition remained stable.

ANOVA results showed that there was no significant effect of learning condition or test timing on either perceived effectiveness or affect (all $p$s > .05).

<table>
<thead>
<tr>
<th>Subjective measure</th>
<th>Interactive learning</th>
<th>Expository learning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Immediate test</td>
<td>Delayed test</td>
</tr>
<tr>
<td>Perceived effectiveness</td>
<td>3.60 (0.89)</td>
<td>3.65 (0.81)</td>
</tr>
<tr>
<td>Affect</td>
<td>3.05 (0.73)</td>
<td>3.06 (0.76)</td>
</tr>
</tbody>
</table>

Note. Values in parentheses denote standard deviations.

4.3 Discussion

Experiment 3 was designed to compare interactive learning with expository learning over an extended period of time. Learner control in the interactive learning condition was defined similar to that in the second experiment. Participants received three worksheets to guide them through the learning process. In comparison, the expository learning condition differed from that in the previous experiment. Because the participants were yoked to those in the interactive learning condition, they experienced
the same simulation scenarios as the interactive learning group. Therefore, the learning content was exactly the same in the two conditions.

A comparison of the interactive learning group's performances on the immediate and delayed test suggested that participants' understanding of the subject matter persisted over time. When tested one week apart, the interactive learning groups had the same level of response accuracy on the knowledge test. They also had the same average statistical reasoning score on the transfer test. As the retention of information is an important criterion of learning, this result provides evidence that the learning these participants gained through interacting with the simulation was of a lasting, rather than transient, nature.

This outcome is important when examined along with the results from the previous experiment. In Experiment 2, interactive learning was found to lead to better test performance than expository learning. However, because participants were tested immediately after the learning stage, it was not clear if the enhanced learning outcome could persist. Schmidt and Bjork (1992) differentiated acquisition performance from retention. They maintained that methods used to maximize performance during learning can be detrimental in the long run, and conversely, methods that degrade the speed of acquisition can support the long-term goals of learning. Therefore, acquisition performance is an imperfect indicator of learning. Schmidt and Bjork's assertion was supported by a variety of experiments on motor and verbal learning. The implication is that only long-lasting performance changes can qualify as learning effect. Because the interactive learning group in the current experiment performed at the same level over the one-week delay, it lends support to the validity of the conclusion in Experiment 2. In
view of the current result, it is reasonable to argue that the enhanced performance in that
experiment reflected the true merit of interactive learning.

Although test performances in the interactive learning condition remained stable
over time, when they were compared to the performances of the yoked group, the current
experiment did not yield support for the advantage of interactive learning over expository
learning. On the immediate test, participants in the two learning conditions had
essentially the same performance on both the knowledge and transfer tests. On the
delayed test, participants in the expository learning condition outperformed the
interactive learning group on both tests, although the differences were not statistically
reliable. There was also no evidence of aptitude-treatment interaction in the current
experiment.

There were a number of reasons that could have possibly contributed to this
result. The first was the extra mental effort associated with learner control in comparison
to the yoked expository learning. Although interactive learning in the current
experiment was structured in the same way as in Experiment 2, there was an important
change in the learning procedure. In Experiment 2, participants were given 20 minutes to
interact with the simulation, and they spent 10 minutes to determine the effect of sample
size on the sampling distribution, which was the most important part of the subject
matter. In the current experiment, participants had only 12 minutes to work on the
simulation. As the time was evenly divided among the three worksheets, they had a much
reduced amount of time (four minutes) to examine the effect of sample size. Although
participants were able to follow through each worksheet, the reduced time allowance
could have forced them to do so in a rush. As learner control entailed learning to use the
simulation program in addition to learning the content, the time pressure could have a more negative effect on these participants than on those who merely observed the interaction sequence.

This explanation is compatible with the cognitive load theory (Sweller et al., 1998). According to this theory, learning and understanding the instructional material causes intrinsic cognitive load while conducting other activities imposed by a specific learning method produces extraneous cognitive load. Inappropriate or excessive learner activities could increase extraneous cognitive load, leading to reduced capacity to solve problems. Therefore, it is important to direct cognitive resources towards the activities most relevant to learning. In the current experiment, learner control in the simulation-based environment incurred a considerable amount of learner activity. These activities increased demands on learners' cognitive resources. When under time pressure, the extraneous mental effort inevitably caused interference with learning the content.

A second possible explanation for the lack of advantage in interactive learning is related to the learning structure posed in each learning condition. In the current experiment, the packet of three worksheets was provided to the interactive learning as well as expository learning groups. The worksheets served to divide the learning process into manageable steps and directed the learner's attention to specific aspects of the simulation. However, while this was necessary and important for learner control in interactive learning, it could also encourage the expository learning participants to be actively involved in the learning process. For example, based on the worksheets they received, participants could generate hypotheses and, without all the burden of conducting physical activities, use the simulation demonstration to verify their
hypotheses. As long as this group remained cognitively active, the difference between interactive and expository learning could have been greatly reduced.

Another reason that could have contributed to the current result is the yoked experimental design. Because participants in the expository learning condition observed the other group's interaction with the simulation program, they were exposed to the increased variability normally would be achieved only with learner control. In comparison, in Experiment 2, although a simulation demonstration was also used in expository learning, it followed a predefined simulation sequence and so was much more restricted than that in interactive learning. As practice variability has been shown to enhance learning and transfer, many attribute the advantage of a simulation to the increased variability a learner experiences in such an environment (Alessi & Trollip, 2001). Viewed from this perspective, it is not surprising that, once the expository learning group experienced the same amount of simulation runs as did the interactive learning group, the difference between the two groups disappeared.

In the current experiment, participants' subjective responses to the learning conditions did not show differentiated effects associated with learner control. Overall, the interactive learning group and the expository learning group had similar positive opinions regarding the effectiveness of learning. They also had moderate affect towards each learning condition. However, the two groups did exhibit different responses over time. Responses of the expository learning group fluctuated with the passing of time. Although their affect towards the expository learning method increased one week after the initial learning, their judgment of its effectiveness became less positive. This response was somehow inconsistent with this group's actual performance, which was the highest
among all the four experimental conditions on both tests. In comparison, participants who interacted with the simulation were more consistent in their attitudes over time. Their ratings on the perceived effectiveness and the affective measure remained stable over the one-week delay. It seemed that exerting control over the learning process allowed these learners a better, more stable ground to base their judgments and attitudes on.

To summarize, the current results showed that simulation-based interactive learning generated stable learning outcome over time. The current experiment did not provide support to the advantage of learner control over expository learning. Comparing the two learning conditions suggested that an optimal combination of learning parameters is needed for learner control to be effective.

5. GENERAL DISCUSSION

The three experiments in the current research investigated the effect of learner control in the context of using a computer simulation to learn a complex statistical principle, the law of large numbers. Interactive learning was structured in different ways in these experiments, and was compared to textbook- or simulation-based expository learning in different experiments.

The cumulative results from the three experiments suggest that learner control has the potential to facilitate learning of complex statistical concepts and lead to better learning outcome compared to expository learning. Specifically, the results showed that, under certain circumstances, interactive learning increased learners’ initial understanding of the subject matter and improved their ability to apply statistical knowledge to real-
world problems. Learner control also had a positive impact on learners’ motivation. However, these results should be interpreted with caution, because the effect of learner control depended on a combination of learning parameters, and the results were not clear-cut in all cases.

There is some suggestion from the current research that the beneficial effect of interactive learning is more evident when learning is measured in terms of learners' ability to apply knowledge directly to solving statistical problems. In Experiments 1 and 2, participants in the interactive learning conditions had higher levels of response accuracy in the statistical knowledge test than participants in the expository learning condition, even though their performances on the transfer test generally did not show such an advantage. It was possible that the increased variability in a simulation-based environment, that is, the varied, vivid learning experience these participants gained through controlling the simulation setting and observing the immediate outcome had helped them achieve a better, more comprehensive understanding of the learning content. In comparison, participants in the expository learning conditions might not attend as closely to the material throughout the learning stage. Roth and colleagues (Roth, McRobbie, Lucas, & Boutonne, 1997) made a similar argument when they observed that students in a high school physics class failed to learn from teaching demonstrations. Their analyses revealed that a number of problems had prevented the students from learning what the demonstration was intended for. One specific problem was the difficulty learners had in separating signals from noise in the demonstration. They also lacked opportunities to test their descriptions and explanations due to the expository nature of demonstration. By contrast, interacting with a simulation involves an active process of
testing hypotheses and interpreting data from experiments. It also requires the learner to
pay close attention to the contingency between variables, forcing them to develop and
refine their representations of the underlying principle.

In the current study, even though learner control did not have an overall beneficial
effect on participants' performance on the transfer test, there was evidence that, under
certain conditions, interactive learning could also improve learners' ability to transfer
statistical knowledge to novel situations. In Experiment 2, it was found that interactive
learning had different results for learners with different levels of cognitive ability.
Specifically, interactive learning significantly improved participants' transfer test
performance for those with medium low cognitive ability. This advantage was much
smaller for the low cognitive ability group, and even diminished for learners with
medium high or high cognitive ability. Although statistically not significant, participants
performance on the knowledge test exhibited the same pattern, with the medium low
cognitive ability group gained the most from interactive learning. Thus, the current result
corroborated previous research finding that, in the specific domain of learning statistics,
interacting with a simulation could be more beneficial for learners with lower, rather than
higher, ability (Veenman & Elshout, 1995; Weir et al., 1991). Although interaction
between cognitive ability and learning condition was not replicated in Experiment 3, it
was likely caused by changes in the learning procedure. The limited time participants had
interacting with the simulation in Experiment 3 could have a negative effect on learning
when compared to expository learning. This disadvantage could have been particularly
detrimental to those with relatively low cognitive ability.
Aptitude-treatment interaction has significant practical implications in education. If there is valid interaction between aptitudes and treatments, instructional methods can be adapted to these individual characteristics to capitalize on assets, compensate for weaknesses, or remedy shortcomings (Cronbach & Snow, 1977). The accumulating body of evidence regarding the selective effectiveness of statistical simulations on different ability learners shows that a particular learning method is preferable to another for a certain type of students. This potentially can be used for decision making in real educational settings.

Results from the current research also show that learner control had positive impact on learners' subjective responses to learning. In the first experiment, learner control resulted in more positive affect towards learning than expository learning. By contrast, in Experiment 2, when the same content delivery medium was used in both learning conditions, interacting with the simulation did not have such an effect on learners' affective responses. This result suggests that some of the differences between interactive and expository learning on the motivational measure could be caused by participants' preference of computer-based delivery medium over traditional print medium. However in Experiment 2, even when the same type of medium was used in both learning conditions, learners still perceived interactive learning to be more effective than expository learning. The results from Experiment 3 further suggest that interacting with the simulation allowed learners to form a better ground to base their judgments on, so that their subjective responses remained stable over time. By contrast, the attitudes of those in expository learning fluctuated during the same period of time.
Motivation was measured as a dependent variable in the current study. In a regular educational setting, however, students' high motivation also has the potential to contribute to a better learning outcome. Motivated learners are more likely to continue thinking about the material after the learning stage and test the hypothesis in domains other than what was encountered during learning (Kersh, 1964).

Although the current research has shown that learner control has the potential to produce better learning outcomes, the cumulative findings here strongly indicate that interactive learning requires a well-defined structure in order to be effective. In the three experiments, interactive learning was structured in three different ways. In the nondirective guidance condition, a set of high-level, nondirective questions was provided to guide participants in the learning process. In the second condition, participants were provided with worksheets. The worksheets divided the learning process into discrete steps and provided assignments during each step. In the third, directive guidance condition, participants were provided with step-by-step instructions on how to interact with the simulation. Thus, interactive learning was increasingly structured in these three conditions. Overall, it was found that learners benefited most from interacting with the simulation when they had the worksheets to guide their interaction. As the structure provided by the worksheets resides somewhere in the middle of the directive-nondirective guidance continuum, the result from the current research is consistent with findings from other empirical research, that structuring the learning process at a moderate level with balanced guidance and exploration is a promising approach to realizing the potential of learner control (e.g., de Jong & van Joolingen, 1998; delMas et al., 1999; Linn & Songer, 1991).
The importance of structure in simulation-based interactive learning has been reflected in a body of research on the closely related topic of discovery learning. Mayer (2004) reviewed research literature on discovery learning spanning over three decades, and drew the conclusion that, across a variety of subject domains, guided discovery was more effective than pure discovery in helping students learn and transfer. Mayer differentiated between two types of learner activities in discovery learning: behavioral activity and cognitive activity, and further argued that it is cognitive activity, rather than behavioral activity per se, that truly promotes learning. Meaningful learning occurs when the learner is engaged in cognitive processes (e.g., selecting, organizing, and integrating information). The often-overstressed behavioral activity in pure discovery does not translate into cognitive processing. Therefore, instructional support is necessary to focus on specified educational goals and guide learner's cognitive activity. Mayer noticed that the challenge of guided discovery is to determine the right type and amount of guidance to provide in discovery learning. This theme, as can be seen, is consistent with the research question in the current work.

In the current research, learner control was found to lead to better learning outcome than expository learning. This advantage, however, was found to depend on other factors. Specifically, the interactive learning condition in Experiment 3 was structured in the same way as in Experiment 2 but was not better than the expository learning condition. This suggests that the effect of interactive learning could be influenced by a number of learning parameters.

A possible explanation of the different outcomes in Experiments 2 and 3 is provided by the cognitive load theory. Viewed from this perspective, learner control in
the interactive learning environment was accompanied with additional demand on learner’s cognitive resources. Compared to expository learning, participants in interactive learning needed to understand the simulation setup before they could focus on learning the content, a task that was quite challenging considering that they had no prior knowledge of the subject matter. As Tuovinen and Sweller’s study (1999) has shown, the structure of learning process has a direct impact on extraneous cognitive load, and the increased mental load is associated with less effective learning results, especially when the learner is not familiar with the subject domain. Other empirical research has provided further support to such analysis. For example, Paas (1992) compared the different effects on test performance and cognitive load of three strategies for learning statistics. The three strategies emphasized solving conventional problems, studying worked-out problems, or completing partially worked-out problems, respectively. The results showed that learning with partially or completely worked-out problems led to less effort-demanding and better transfer performance. In view of the results from the current research, the extraneous cognitive load associated with learner control could have interfered with learning under certain circumstances, such as when the learning structure was less than ideal or when there was time pressure. Therefore, it is important to recognize the constraints of learner control and use caution when designing instructional guidance for interactive learning.

In summary, the current research generated valuable information regarding learner control in a simulation-based environment. Two findings, those of the importance of learning structure and the differentiated effects of learner control on learners with different cognitive abilities, corroborated with previous research and have important practical implications. However, in view of the richness of this research topic, the current
work may be furthered in a number of directions. First, the three experiments compared
different learning structures in interactive learning. As de Jong and van Joolingen (1998)
pointed out, structuring the discovery process is an instructional measure designed to aid
the regulative process during interactive learning. Researchers have started looking at
providing learning support for other processes during scientific discovery learning, such
as hypothesis generation, experiment design, and data interpretation, but empirical
research on these aspects is scarce. Therefore, it is worthwhile to compare learning
support along other dimensions as well. Second, it will also be of value to use simulation-
based environment to study transfer. Despite of a century s worth of research,
psychologists still differ in their opinions as to whether far transfer occurs. For those who
believe that far transfer does occur, its underlying mechanism remains an open question.
In a simulation-based learning environment, learning is, to a large extent, under the
control of specific environment parameters (Sedlmeier, 1999). Therefore, computer-
based simulations provide a good platform for investigating what elements of the learning
process might be responsible for effective transfer. Third, as experiments conducted in a
laboratory setting, the current research is limited in its ecological validity. Thus, it will be
of both theoretical and practical value to conduct similar research in a regular educational
setting.
PART II: EVALUATING THE USABILITY
OF A WEB-BASED LEARNING ENVIRONMENT

Part I of this thesis is focused on researching the learning outcomes of using an interactive computer simulation in learning statistics. In this part I turn to another type of interactive learning environment, that is, Web-based learning, and discuss two usability studies that aimed to evaluate and improve learner control in such an environment. Like a simulation, a Web-based learning environment encourages learners' active participation by giving them control over the learning process. However, while the scope of a computer simulation is usually limited to one specific topic within a subject area, a Web-based learning environment typically encompasses the whole curriculum in a subject domain. As a result, learner control in a Web-based environment is multi-faceted, complex, and has much of an impact on the learning process. These distinct characteristics have promoted new approaches to investigating learner control in Web-based learning.

1. INTRODUCTION

Learner control in Web-based learning

The rapid advancement in information technology has made the World Wide Web an increasingly important part of the infrastructure of the educational system. A Web-based learning environment is essentially a hypermedia system that uses the attributes and resources of the Web to facilitate learning. Many educators consider Web-based
learning to be of a constructivist nature, because it is learner controlled, learner modifiable, and it supports multiple ways of navigation and information search.

Two unique features of Web-based learning are particularly relevant to the much increased learner control that differentiates Web-based learning from traditional ways of learning. First, information on the Web is organized in a non-linear fashion, and is easily accessible. Because of this, the learner has great flexibility in choosing the content, selecting the path, and setting the pace during the learning process. Consequently, Web-based learning takes on a self-directed characteristic, with the learner in control of much of his/her own learning experience. Second, the Web is a powerful delivery system that integrates multimedia as well as multiple learning methodologies. A Web-based environment not only provides a vast amount of textual, visual and auditory information, it also accommodates computer-based learning methodologies that involve a wide range of learner interactivity, from tutorial and drill to simulation and educational game. Therefore, learner control in such an environment goes beyond sequencing and is achieved at the more fundamental content level.

Given the complexity of a Web-based system, there is yet another aspect to learner control in Web-based learning, that is, the human-computer interaction. This aspect concerns the physical process of learner interactivity and emphasizes the compatibility between the learner and the interface design. Viewed from this perspective, a learner essentially becomes a user of the learning system, and learner control is frequently referred to as user interaction.

User interaction is fundamental and important in Web-based learning, because it is through interacting with the learning environment does a user learn its structure and
subsequently exert learner control at the pedagogical level. It is at this level that learner control will be examined later in this part.

**Usability and Web-based learning environment**

The enhanced learner control in a Web-based learning environment requires that the system is carefully designed to provide good usability. From the usability perspective, a good computer system is one that is easy to learn, efficient to use, easy to remember, keeps the error rate at the minimal, and maintains a high level of user satisfaction (Nielsen, 1993). The same standards apply to a Web-based learning environment.

A number of issues are critical in determining the usability of a Web-based learning environment. The most prominent usability problem associated with Web-based learning is navigation. Due to the Web's non-linear style of information access, a learner may easily get confused about his/her whereabouts in the information space and become disoriented (Stemler, 1997). Another issue that is of great importance to Web-based learning is screen design. The optimal use of visual elements, such as color, text, graphics and animation plays a crucial role in the delivery of instructional materials on a computer display (Grabinger, 1996). In general, a Web-based learning environment is usable only when it is comprehensive, visually consistent, easily navigated and organized around the tasks that the learners intend.

There are different approaches to improving the usability of a Web-based environment, including guidelines, heuristic evaluation, user testing, and so on (Nielsen, 1993). These different methods have a varied degree of success in practical settings. In particular, research has shown that empirical user testing is one of the most effective
ways to uncover usability problems. In a user testing, a small number of representative users use the software to perform a set of tasks, and their interaction with the system is carefully monitored and analyzed to provide detailed information as to whether certain design features have achieved the intended purpose. The results from a user testing may be used to guide the software iterative design process. Researchers have reported positive experiences applying this method to improving the usability of Web-based learning environments (e.g., Chisman, Walbridge, & Diller, 1999).

The Connexions Web-based learning environment

The Connexions project was a Web-based courseware designed for the collaborative creation and dynamic delivery of high-quality educational materials. The learning environment was built upon two key concepts. The first was the modularization of knowledge. Knowledge modules were encapsulated packages of information that cover specific topics in a subject area and contained additional pedagogically-related contents, such as prerequisites, applications, multimedia, and supplementary materials. The second key concept was Web-based navigational aids that allowed learners to explore the inter-connectedness among various modules (Hendricks, 2001). In the Connexions learning environment, modules were created as individual building blocks and subsequently assembled together to form courses, and navigational tools were provided in the student user interface to help learners view course materials online.

From a usability perspective, there were three issues most relevant to user interaction in the Connexions learning environment. First and foremost, the hypermedia methodology used for content organization and presentation required sophisticated
navigational aids to keep learners oriented in an intricately-woven web of information. The second issue had to do with icon recognition. The student user interface used various icons to represent different system functionalities, some of which were unique to the Connexions learning environment. For example, it devised the concept of "Link Strength" to describe the degree of inter-connectedness between pieces of information. It remained a question if this novel concept was valid for pedagogical purpose and, if so, what would be the best way to visualize this concept. Third, because exerting control in the Connexions environment required a learner to possess procedural knowledge such as setting up the system in a Web browser or loading a course, it was very important that the information was easy to find and easy to understand. In summary, navigation, icon recognition, and information search were considered the most critical user interaction in the Connexions learning environment. As a result, usability evaluation of the student user interface was focused exclusively on these three aspects.

The purpose of the work presented in part II was two-fold. First, it was conducted to empirically evaluate the usability of the Connexions learning environment and use the information gathered to aid software development. Second, it aimed to look into the strength and limitations of user testing as an empirical usability evaluation method. This goal was accomplished by comparing results from two consecutive iterations of user testing.
2. USER TESTING 1

2.1 Method

2.1.1 Participants

Six Rice University undergraduate students (all engineering majors) volunteered in the testing. On average, participants were 20 years of age, and in their third year of study. All participants had moderate (one to three years) to extensive (more than three years) Web experience, but had very limited experience using online textbooks. All had normal or corrected-to-normal eyesight.

2.1.2 Materials

The student user interface of the Connexions learning environment is shown in Figure 11.
Figure 11. Screenshot of the initial Connexions student user interface. The interface was divided into three functional areas. The left panel displayed the organization of modules in a course. Modules were typically organized in a linear fashion, but allowed random access. The middle panel displayed the content of a selected module. It also had a number of icons on the top that served navigational and other functions. The right panel displayed the links related to the module. On its top were the icons representing link strength. In this initial design, each icon corresponded to one particular level of link strength.

Eighteen tasks were used to tap into the usability of the Connexions student user interface. These tasks were typical of learners’ interaction with the learning environment. Among them, eight tasks were related to navigation, six were about icon recognition, and the remaining four dealt with information search. The complete task list is provided in Appendix G.
2.1.3 Procedure

Participants were tested individually. Upon arriving in the lab, each participant first provided information on his/her Web and online learning experience by filling out a short questionnaire. Then the participant interacted with the user interface to perform the tasks. During this process, the participant was asked to speak out loud his/her thoughts, and the whole testing session was videotaped. In the meanwhile, two data collectors who were familiar with the user interface observed from behind and took notes of the participant’s performance independently. After finishing the tasks, the participant filled out a second questionnaire giving comments on the user interface. The testing session lasted less than one and half hours for all participants.

2.1.4 Data Analysis

Notes taken by the two data collectors were first compared and any inconsistency was reconciled by referring back to the video recording. The notes recorded how participants interacted with the user interface on each task. The performance was then compared to the prescribed interaction sequence. If there was a mismatch, the performance was counted as a failure, and the participant’s vocal protocol was analyzed to provide more detailed information on this particular task. Participants’ comments on the user interface were used to corroborate findings from performance analysis.

2.2 Results and Discussion

Users’ task performances in the first user testing are summarized in Table 4. Overall, the six users failed on the tasks for a total of 36 times. While some tasks were
fairly easy for all the users, others were more problematic. In particular, there were eight tasks on which more than half of the users failed to perform. These most difficult tasks were distributed in all three task categories (three in navigation, three in icon recognition, and two in information search).

Table 4. Observation of user performances in the first user testing

<table>
<thead>
<tr>
<th>Task</th>
<th>Number of Errors</th>
<th>Errors Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 6</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Task 7</td>
<td>3</td>
<td>failed to use the browser back button</td>
</tr>
<tr>
<td>Task 8</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Task 9</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Task 12</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Task 13</td>
<td>4</td>
<td>ignored the highlight on the left</td>
</tr>
<tr>
<td>Task 14</td>
<td>3</td>
<td>misunderstood Previous &amp; Next</td>
</tr>
<tr>
<td>Task 15</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Icon recognition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Task 3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Task 5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Task 10</td>
<td>3</td>
<td>failed to understand Link Strength icons</td>
</tr>
<tr>
<td>Task 11</td>
<td>3</td>
<td>selected one icon to represent a range</td>
</tr>
<tr>
<td>Task 18</td>
<td>3</td>
<td>failed to find the icon</td>
</tr>
<tr>
<td>Information search</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 1</td>
<td>4</td>
<td>failed to follow the right link</td>
</tr>
<tr>
<td>Task 4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Task 16</td>
<td>6</td>
<td>tried to load a course through the browser</td>
</tr>
<tr>
<td>Task 17</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
<td></td>
</tr>
</tbody>
</table>
Many of the most difficult tasks addressed design issues that were unique to the Connexions learning environment and also critical to user interaction. For example, Tasks 10 and 11 dealt with the visualization of the concept of link strength. The errors observed from users' performances on these two tasks indicated that the mapping scheme was incongruent with users' intuition. Whereas the design stipulated that each individual icon corresponded to a specific level of link strength, users tended to think of each icon as representing a range (from the designated level to the maximum) in link strength. Similar to this example, the other difficulties users experienced during testing also indicated problems with the interface design.

By looking into users' performances on various tasks in the learning environment, the first user testing identified various usability problems with the user interface design. Subsequently, tentative solutions were generated for these identified problems and provided to the software designers. Many of the suggestions were promptly adopted in the new version of the student user interface (for details, see Figure 12 in the next section).

Experts in usability engineering advocate conducting iterative usability testing during software design (e.g., Nielsen, 1993). There were several reasons that supported this practice in building the Connexions Web-based learning environment. First, the interface changes intended to address currently identified usability problems might not necessarily solve these problems, and therefore should be verified through further empirical testing. Second, it was possible that additional usability problems could emerge in iterative testing after the most severe problems had been corrected. Third, because all the six users in the first testing had homogeneous academic background, it raised concern
if their performances were representative of the typical users of the Connexions learning environment. Based on these considerations, a second user testing was conducted on the newly improved user interface.

3. USER TESTING 2

3.1 Method

3.1.1 Participants

Six undergraduate students from four different universities were recruited to participate in this study. Participants had diverse academic background (three majored in humanities, two in engineering, and one in social sciences). Their other characteristics were similar to those in the first user testing.

3.1.2 Materials

Figure 12 shows the user interface in the second user testing.
Figure 12. The improved student user interface based on results from the first user testing. Compared to the previous interface design, major changes were made on the right-hand panel as well as on the top portion of the middle panel. On the right, the original triangle-shaped link strength icons were replaced with the more conventional rectangular design. The slider bar on the top allowed users to define the link strength level at which the links were to be displayed. The functional buttons, originally on the top of the middle panel, were replaced with graphical icons and moved to the top of the whole window. The two navigation buttons ↓ and ↑, which were largely ignored by the users in the first testing, were eliminated.

Nineteen tasks were used in the second testing. The first 18 tasks were the same as in the previous user testing, except that Task 10 was modified to reflect changes in the interface design. In addition, Task 19 was added to examine if users were familiar with the Web browser’s built-in function. The task fell into the category of information search. The complete task list is included in Appendix G.

3.1.3 Procedure and Data Analysis

Same as in the first user testing.
3.2 Results and Discussion

Table 5 summarizes six users’ performances on the 19 tasks in the second usability testing. In order to compare with the first user testing, performance data on the original 18 tasks are discussed first below.

Table 5. Observation of user performances in user testing 2

<table>
<thead>
<tr>
<th>Task</th>
<th>Number of Errors</th>
<th>Errors Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 6</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Task 7</td>
<td>4</td>
<td>failed to use the browser back button</td>
</tr>
<tr>
<td>Task 8</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Task 9</td>
<td>3</td>
<td>misunderstood the link Previous</td>
</tr>
<tr>
<td>Task 12</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Task 13</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Task 14</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Task 15</td>
<td>3</td>
<td>misunderstood the link Previous</td>
</tr>
<tr>
<td>Icon recognition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Task 3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Task 5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Task 10</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Task 11</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Task 18</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Information search</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Task 4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Task 16</td>
<td>6</td>
<td>tried to load a course through the browser</td>
</tr>
<tr>
<td>Task 17</td>
<td>5</td>
<td>failed to find the link Send suggestion</td>
</tr>
<tr>
<td>Task 19 (new)</td>
<td>5</td>
<td>failed to locate browser function</td>
</tr>
<tr>
<td>Total</td>
<td>39</td>
<td></td>
</tr>
</tbody>
</table>
On the first 18 tasks, the six users failed to complete a task for a total of 34 times. Five tasks appeared to be the most problematic, causing more than half of the users to fail. Among them, three tasks were related to navigating in the learning environment, and the remaining two had to do with information search.

A comparison of the most difficult tasks in the two rounds of user testing is provided in Table 6. Across the two testing, the biggest performance improvement occurred on icon recognition. Whereas in the first testing three tasks in this category posed severe problems, none remained so in the second testing. Such performance improvement was directly associated with improvement in interface design. For example, the new representation of link strength, a slider bar, fit users' mental model of this concept. As a result, all users succeeded on Tasks 10 and 11.

<table>
<thead>
<tr>
<th>Task category</th>
<th>First user testing</th>
<th>Second user testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigation</td>
<td>7, 13, 14</td>
<td>7, 9, 15</td>
</tr>
<tr>
<td>Icon recognition</td>
<td>10, 11, 18</td>
<td>None</td>
</tr>
<tr>
<td>Information search</td>
<td>1, 16</td>
<td>16, 17</td>
</tr>
<tr>
<td>Total number of difficult tasks</td>
<td>8</td>
<td>5</td>
</tr>
</tbody>
</table>

Compared to icon recognition, users' performances in navigation and information search categories showed more mixed results. Users did make major progress on three tasks (Tasks 13, 14 and 1), but their poor performances on Tasks 7 and 16 remained the same. Moreover, there were two additional tasks identified as being difficult (Tasks 9 and 15). The persistence of certain usability problems indicated that the relevant design changes fell short of solving the original problems. For example, all six users in the
second testing failed on Task 16 (i.e., loading a course according to its URL), despite the
fact that the entry box for typing course URL had been moved to a more prominent
location in the new user interface than in the old one. On the other hand, the newly
identified problems with the two additional navigation tasks suggested that they might
have been overlooked in the first testing, as no specific interface changes had been made
in that respect.

On the 19th task, five out of six users had trouble locating the web browser’s built-in function (i.e., increasing font size), contrary to what had been expected. As this task
was added to address a future design topic, this result caused the software designers to re-
consider their options.

The results from the second user testing again served as input for the iterative
software development process. Efforts were made to address the identified usability
problems. This is reflected in Figure 13, which shows the most recent student user
interface design.
Figure 13. The improved student user interface based on results from the second user testing. Based on the findings from the second testing, the layout of the module content was modified by putting module summary first while moving extraneous author/maintainer information to the bottom of a module. This change was expected to serve important pedagogical function.

4. GENERAL DISCUSSION

An important aspect of learner control in a Web-based learning environment is for a user to effectively and efficiently interact with the system. Good usability is achieved by matching software design with the perceptual, cognitive, and motor capabilities and constraints of the learner while taking into account the characteristics of the tasks learners perform. In the current study, usability evaluation was conducted to assess the user interface of the Connexions Web-based learning environment. The evaluation examined users' task performances in navigation, icon recognition, and information search, and
revealed design problems in each area. The information served to guide the iterative software design process. As users' task performances in the two consecutive testing showed, there was considerable usability improvement in certain areas across the two interface designs. Most noticeably, the visualization of the link strength concept was directly impacted by the user testing result. Thus, the present work demonstrates that usability evaluation is instrumental to ensuring effective learner interaction in a complex Web-based learning environment, such as the Connexions system.

Furthermore, the current study demonstrates the utility of user testing in software usability engineering. As a matter of fact, when each version of the user interface was developed, it first underwent an informal usability analysis through heuristic evaluation. However, it was not until the testing users performed specific tasks and experienced difficulty did it become clear that some quite significant usability problems had been overlooked. The initial design of the link strength concept best exemplified this situation. This finding is consistent with the usability research literature, wherein it has been reported that, among a number of usability inspection methods, user testing is particularly effective in uncovering the most serious usability problems (Jeffries, Miller, Wharton, & Uyeda, 1991).

User testing should be conducted in several iterations in order to provide an accurate account of the usability of a system and to assess the software revision effort. As demonstrated here, while a single round of user testing can uncover a large number of usability problems, it may also fail to detect some others. Besides, interface changes designed to correct the problems may not work the way they are intended. Therefore, conducting iterative user testing should be considered an integral part of the software
design process. This is all the more important for building an advanced interactive learning environment, because learning happens only when learners are able to effectively interact with the environment in the first place.

User interaction is an important aspect of learner control in Web-based learning. Although usability is studied in isolation in the current research, it is intricately related to pedagogy in a genuine learning environment. Research has shown that an individual’s self-directed learning process in an interactive environment is directly influenced by the usability of the system. For example, in an experiment conducted by Burke, Etnier, and Sullivan (1998), a group of fifth graders learned the solar system with a hypermedia program. The effect of usability on learner control was examined by providing navigational aids to some of the students but not the others. The results showed that those with navigational aids deviated from a linear path significantly more than students who did not have navigational aids. In addition, these students also performed better on the posttest than the latter group. These researchers reasoned that navigational aids might help provide a structure that promoted more exploratory behavior in the students, leading to the superior learning outcome.

In another study, Parlangeli, Marchigiani, and Bagnara (1999) examined the effects of usability on learning more directly. They first conducted a comprehensive usability evaluation of a hypermedia course. Both heuristic evaluation and empirical user testing revealed that the system had various problems in user interface design. After determining its usability, the researchers conducted an experiment to assess the learning outcome. Three groups of college students learned economics using either the hypermedia course, the printed version of the course, or a book. Although students in the
hypermedia course were the only ones who had the convenience to freely explore the contents, their performances after the learning were not different from the performances in the other two groups. Because these researchers expected superiority of the hypermedia learning tool over either the printed version or the book, this negative finding on learning outcome was attributed to the low level of usability of the interactive course.

As demonstrated in empirical research, user interaction is an integral part of learner control in a Web-based learning environment. It has a significant impact on the learning process as well as the learning outcome. Therefore, although the work reported here focused exclusively on this aspect, it should be extended in the future to embrace the full notion of learner control in a complex interactive learning environment.
CONCLUSIONS

Learner control is a valued concept in both the cognitive and constructivist learning paradigms. The rapid advancement of information technology has made learner control an essential component of a student’s learning experience. The much enhanced learner control, in turn, has required more vigorous efforts be devoted to researching its impact on the learner and the learning process.

The three experiments in Part I compared simulation-based interactive learning with expository learning. The cumulative results demonstrate the beneficial effect associated with learner control but stress the importance of learning structure in interactive learning. In addition, it was also found that learner characteristics interacted with learner control such that learners with lower cognitive ability benefited more from interactive learning than those with higher cognitive ability. This finding corroborates previous research in the same area, and provides further empirical evidence for the existence of aptitude-treatment interaction, a long-researched topic in psychology and education.

Learner control in Web-based learning is more complicated and thus puts high demand on the usability of the interactive learning environment. The usability assessment of the Connexions learning environment provided valuable information for the iterative design of the software. It also demonstrates the utility of user testing in this process.

Although the values of learner control were examined in a simulation-based learning environment, the research approach taken in the current work and the results obtained have implications for larger-scale interactive learning environments as well. Web-based learning and simulation-based learning share the characteristic of learner
control. Research on Web-based learning has repeatedly addressed the same research
topics as in simulation-based learning, such as learning structure and aptitude-treatment
interaction. Converging evidence from carefully designed and well-controlled empirical
research should shed light on the effective use of a Web-based learning environment.
REFERENCES


APPENDIX A: EXCERPTS OF DIRECTIVE AND NONDIRECTIVE GUIDANCE FOR INTERACTIVE LEARNING IN EXPERIMENT 1

Excerpt of step-by-step instruction as directive guidance:

Sample size, Uniform parent distribution
1. Set the parent distribution to "Uniform" distribution.
2. To the right of the third graph, set the sample size to "N=2".
3. To the right of the bottom graph, select "N=10".
4. Click the button "Animated Sample" 5 times. Note how the animated processes are related to each of the two bottom graphs.
5. Click on the button "5000 Samples" 10 times. Compare the mean, the standard deviation, and the overall shape of the two offspring distributions.
6. To the right of the third graph, set the sample size to "N=10".
7. To the right of the bottom graph, select "N=25".
8. Click on the button "5000 Samples" 10 times. Compare the mean, the standard deviation, and the overall shape of the two offspring distributions.
9. To the right of the third graph, select sample size to be "N=2".
10. To the right of the bottom graph, select "N=25".
11. Click on the button "5000 Samples" 10 times. Compare the mean, the standard deviation, and the overall shape of the two offspring distributions.

Excerpt of high-level questions as nondirective guidance:

1. How does the sample size influence the mean and the standard deviation of the offspring distribution? Note that the sample sizes of the two bottom graphs can be set at different values so you can see the two sampling distributions at the same time.
2. Move the vertical bars on the bottom two graphs along their respective X-axis but always at the same position. Compare the areas to the left (or right) of the bars in the two graphs.
3. If you find that sample size influences the offspring distribution, does that relationship hold true for parent distributions of different shapes?
APPENDIX B: MULTIPLE-CHOICE QUESTIONS USED IN EXPERIMENTS 1, 2 AND 3

1. Standard deviation is a statistical term frequently used to describe ________.
   (a) the arithmetic average of all the data points in a distribution
   (b) the randomness by which samples are drawn from a population
   (c) the extend to which each individual scores in a distribution cluster together
   (d) how many times does the sample mean differ from the population mean

2. The shape of the sampling distribution of the mean based on sampling from a normal population distribution many times ________.
   (a) is a uniform distribution
   (b) is a normal distribution
   (c) is a skewed distribution
   (d) depends on the sampling process

3. If we draw a large number of samples of size 25 from a skewed parent distribution, the shape of the distribution of the sample mean will resemble ________.
   (a) a uniform distribution
   (b) a normal distribution
   (c) the parent distribution
   (d) depends on the sampling process

4. If we draw a large number of samples of size 10 out of the population, the mean of the sampling distribution ________ the mean of the population.
   (a) is greater than
   (b) is less than
   (c) is equal to
   (d) depends on the raw data

5. If we draw a large number of samples of size 10 out of the population, the standard deviation of the sampling distribution of the mean ________ the standard deviation of the parent population.
   (a) is greater than
   (b) is less than
   (c) is equal to
   (d) depends on the raw data

6. The standard deviation of the sampling distribution of the mean for samples of size N=5 ________ that of the sampling distribution of the mean for samples of size N=10 (assuming both sampling distributions include a large number of samples).
   (a) is greater than
   (b) is less than
   (c) is equal to
   (d) cannot be determined from the information given
7. Which of the following said about a normal distribution is true? ______.
   (a) A normal distribution is the data obtained from a group of normal people
   (b) A normal distribution has several peak values
   (c) A normal distribution is a mathematically normalized dataset
   (d) A normal distribution resembles a bell shape

8. Suppose we have a normal distribution with a mean of 30 and a standard deviation of 8. M is the number of data points that fall below 28, and N is the number of data points that fall beyond 42, and N is the number of data points that fall below 28, then ______.
   (a) M > N
   (b) M = N
   (c) M < N
   (d) which one has more cases cannot be decided based on the information above.

9. Suppose we have a population distribution with a mean of 100 and a standard deviation of 15. We randomly draw a sample of 2 and another sample of 5 from the population. Which sample is more likely to have a sample mean smaller than 85?
   (a) The one that has a sample size of 2
   (b) The one that has a sample size of 5
   (c) Both samples are equally likely
   (d) Cannot be determined from the information given

10. Which of the following about sample size is true? ______.
    (a) Sample size means how many times samples are drawn from the population
    (b) Sample size means how many cases are included in a single sample
    (c) The bigger the sample size, the bigger the standard deviation of the sampling distribution of the mean
    (d) Once we have information about the population distribution and the sample size, we can predict exactly what the mean value of each individual sample would be.
APPENDIX C: SCENARIOS USED IN THE TRANSFER TEST
IN EXPERIMENTS 1, 2 AND 3

Scenarios 1 through 10 are transfer questions that require applying the law of large numbers. Scenarios 1 through 6 were used in all three experiments. Scenario 7 was used in Experiment 1 only, and Scenarios 8 through 10 were used in Experiments 2 and 3 only. Scenarios 11 through 13 are filler items irrelevant to the law of large numbers. They were used in all three experiments.

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, 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 as more favorable than the one on the left.

Assuming that the old man has nothing to gain from the slot machines, do you agree with Keith on his conclusion? Please explain your answer.

2. 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 Volvos 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 that same day, which do you think they should choose? Justify your answer.

3. 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.

Should the student be admitted? Justify your answer in light of the strengths and weaknesses of the arguments put forth by the two reviewers.

4. 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.

If coach Graves needs to choose the best quarterback he can get and he can only choose one, which quarterback should he choose, Mike or Scott? Justify your answer.

5. There are two hospitals in Normaltown, TX: Central and Valley. At Central Hospital, 100 babies are born each day. At Valley Hospital, 30 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower. For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?

There are only three possible answers to this question. (1) Central hospital; (2) Valley hospital; or (3) About the same (i.e. within 5% of each other). Which one do you think is the correct answer? Please justify your answer.

6. 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. 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.

Which one of the two investors would be more likely to realize a 5% gain over the last 5 years? Please justify your answer.

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. Mark has to play by the same rules Joe chooses.

How do you think Joe maximize his odds of winning? Please justify your reasoning.

8. Mark is a teacher at a small mid-western college. During his first year of teaching, he selected a new textbook for his introductory class, An Introduction to Psychology. He likes this book very much. Later he learns that his colleagues at a much larger university have been using both his book and a different text, Psychology for the
Ages, in their introductory psychology classes. The colleagues report that Psychology for the Ages is much easier for the students to understand and that the students seem to like it better. The students at both colleges have comparable academic background.

Should Mark switch to the other textbook or should he continue to use the one he likes? Explain your answer.

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? Explain your answer.

10. Two sports fans are arguing over which sport—baseball or football—has the best (most accurate) play-off system. Charlie says that the Super Bowl is the best way of determining the world champion because, according to him, "the seven games of the world series are all played in the home cities of the two teams, whereas the Super Bowl is usually played in a neutral city. Since you want all factors not related to the game to be equal for a championship, then the Super Bowl is the better way to determine the world championship."

Which procedure do you think is a better way to determine the world champion—World Series or Super Bowl? Explain your answer.

11. 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 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.

Did Martha have good evidence for feeling she doesn’t like vacations? Please Justify your answer.

12. A brewery buys nearly all of its glass bottles from a local manufacturer. Once the local company is unable to deliver enough bottles and so the brewery orders a shipment from another glass manufacturer that distributes its products nation-wide. On the first day these new bottles are used, the bottle-filling machinery has to be stopped four times because of jamming and so production for the day is unusually low (Ordinarily there is no more than one jamming a day.) The foreman is worried and 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 on 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.

Is the foreman’s reasoning basically sound? Please justify your answer.

13. An economist was arguing in favor of a guaranteed minimum income for everyone. He cited a recent study of several hundred people in the United 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 would 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.

How do you think the economist’s reasoning? Please justify your answer.
APPENDIX D: QUESTIONNAIRE USED IN EXPERIMENTS 1, 2 AND 3

Questions 1 through 8 were used in all three experiments. Questions 9 and 10 were used in Experiment 1 only. Question 9a was used in the textbook-based expository learning condition, and Question 9b was used in both simulation-based interactive learning conditions.

Participants answered each question on the following 5-point scale (the anchor words were modified for certain items to match the specific question):

1-------------2-------------3-------------4-------------5
Not at all    A little    Somewhat    Quite a bit    Very well

1. How well do you understand the basic statistical terms introduced here, such as frequency distribution, standard deviation, sample size, and sample mean?

2. How well do you understand what a sampling distribution is?

3. How well do you understand the relationship between the shape of the population distribution and the shape of the sampling distribution?

4. How well do you understand the relationship between sample size and the standard deviation of the sampling distribution?

5. How much confidence do you have in your answers to the questions?

6. How well do you think the training material helped you answer the questions?

7. How much enjoyable did you feel when learning this specific material?

8. In general how much effort did it take for you to understand what was introduced in the training?

9a. In this experiment there is another group of students who receive the training through interacting with a computer simulation. Was it easier to learn from reading a text chapter than to learn from a simulation?

9b. In this experiment there is another group of students who receive the training through reading a chapter from a textbook. Was it easier to learn through interacting with a computer simulation than to learn from a textbook?

10. If you are asked to participate in another statistical training experiment, and again there will be a computer simulation group and a text group, how desirable is it for you to remain in the current condition?
APPENDIX E: SAMPLE WORKSHEET PROVIDED IN THE INTERACTIVE LEARNING CONDITIONS IN EXPERIMENTS 2 AND 3

Step 1: Focus on the shape of the distribution of sample means.

Question: Do the distributions of sample means resemble their respective population or do they share common feature regardless of the shape of their respective population?

<table>
<thead>
<tr>
<th>Population</th>
<th>Uniform</th>
<th>Normal</th>
<th>Skewed</th>
<th>Custom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Result</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Population</th>
<th>Uniform</th>
<th>Normal</th>
<th>Skewed</th>
<th>Custom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Result</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX F: ANCOVA RESULTS FROM EXPERIMENT 2

Participants' performances on the multiple-choice test and the transfer test were respectively analyzed as ANCOVA with the following model specification:

**Variables**

Learner Control: Two levels. Textbook-based expository learning vs. Simulation-based interactive learning.

Prediction: Two levels. No prediction vs. Prediction.

Covariate: Wonderlic test score

**ANCOVA results of accuracy on the multiple-choice test are shown below:**

<table>
<thead>
<tr>
<th>Effect</th>
<th>F</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive ability</td>
<td>80.5389</td>
<td>1, 103</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Learner control</td>
<td>6.7097</td>
<td>1, 103</td>
<td>0.0110</td>
</tr>
<tr>
<td>Prediction</td>
<td>2.1600</td>
<td>1, 103</td>
<td>0.1447</td>
</tr>
<tr>
<td>Learner control x Prediction</td>
<td>1.3278</td>
<td>1, 103</td>
<td>0.2519</td>
</tr>
<tr>
<td>Learner control x Cognitive ability</td>
<td>2.7056</td>
<td>1, 103</td>
<td>0.1030</td>
</tr>
<tr>
<td>Prediction x Cognitive ability</td>
<td>0.2228</td>
<td>1, 103</td>
<td>0.6379</td>
</tr>
</tbody>
</table>

**ANCOVA results of statistical reasoning on the transfer test are shown below:**

<table>
<thead>
<tr>
<th>Effect</th>
<th>F</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive ability</td>
<td>135.5622</td>
<td>1, 103</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Learner control</td>
<td>0.6721</td>
<td>1, 103</td>
<td>0.4142</td>
</tr>
<tr>
<td>Prediction</td>
<td>0.1196</td>
<td>1, 103</td>
<td>0.7302</td>
</tr>
<tr>
<td>Learner control x Prediction</td>
<td>0.0019</td>
<td>1, 103</td>
<td>0.9652</td>
</tr>
<tr>
<td>Learner control x Cognitive ability</td>
<td>6.0054</td>
<td>1, 103</td>
<td>0.0159</td>
</tr>
<tr>
<td>Prediction x Cognitive ability</td>
<td>1.3602</td>
<td>1, 103</td>
<td>0.2462</td>
</tr>
</tbody>
</table>
APPENDIX G: TASKS USED IN USER TESTING 1 AND 2

Tasks 1 through 18 were used in both user testing. In particular, Task 10a was used in the first testing, and Task 10b was used in the second testing. Task 19 was used only in the second testing.

1. Suppose you are using this web-based textbook environment for one of your classes. You are currently using the web browser Mozilla (version 1.0). Start from http://cnx.rice.edu/browse/, install the Roadmap.

2. Once you install the Roadmap, notice any changes in the browser window, and tell us what function you think these icons might have.

3. The system provides a Roadmap tutorial. Find it and read the content briefly.

4. Load course "Rice University Elec599".

5. Read module "QML 1.0 tutorial" in chapter "other Markup language". Read the "Module Summary" section only.

6. Read the first three modules in the chapter "CNXML". With each module, only read the "Module Summary" to get a brief idea of what that module is about.

7. Read module "The Basic CNXML" in chapter "CNXML", Go down to the "Conclusions" section. Click on the link "The Advanced CNXML", read briefly the "Conclusions" section. Click on the link "The CNXML Language Spec", and read briefly. Go back to the module you visited prior to the current module.

8. In "The Basic CNXML" module, click on link "Connexions Vision". Read briefly the last section "CNXML language". Go back to the module you visited prior to the current page.

9. Select module "Combining XML languages" in chapter "XML". Click on the first link in the content ("XML"). On the new page, click on the first link ("Core Drafts"). Again on the new page click on the 2nd link (XML Inclusions). Read briefly. Then click on link "Copyright". Read briefly. Go back to the last module you visited in the course.

10a. On top of the right panel are icons that represent link strength. Tell us how you think these icons are related to the links displayed below? How do you interpret each individual icon?

10b. Each link on the right panel has a link strength according to its relevance to the module. On top of the panel is a slider that controls the link strength level at which links are displayed. Tell us how you think of the brownish icons preceding each link. How is the slider related to the links appearing below?
11. Select module Combining XML languages in chapter XML.
   - Step 1: Click on link page combining CNXML and MathML in category Example.
   - Step 2: Display only the links of at least medium strength.
   - Step 3: Select module XML basics in chapter XML. Find the link "A page written in XML" under category "Example".

12. Read "Content MathML" in chapter "Other Markup languages". Hide the left panel.
    Click the last link "MathML Specification" and read briefly. Then go to the next module in the course and read the first paragraph.

13. If the left panel is still hidden, restore it. Find a module entitled "Fourier Analysis".
    Read the 1st hit result. Then go back to the module you were reading before you initiated the search.

14. Look up the word "consortium". Then continue reading the module you were on before this task.

15. Find out information about the sponsors of the Connexions project. Then continue reading the module you were on before you initiated the search.

16. Load a course with its URL "http://cnx.rice.edu/courses/elec305/course.rdf". Read the first module in the first chapter. Read only the first paragraph.

17. Submit a suggestion "include more engineering courses" to the design team.

18. Turn off Roadmap to resume normal browser view.

19. Increase the font size of the current display to size = 18.